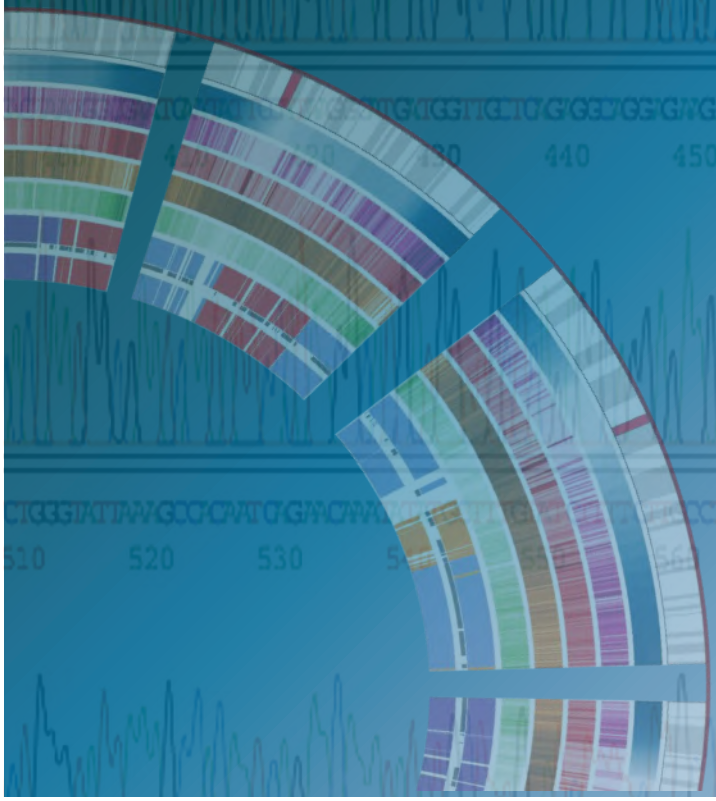
# 8th Annual Sequencing, Finishing, Analysis in the Future Meeting



Santa Fe, New Mexico
May 29-31, 2013

Los Alamos
NATIONAL LABORATORY
EST. 1943

# *Contents*



*The 2013 "Sequencing, Finishing and Analysis in the Future" Organizing Committee:*

** Chris Detter, Ph.D., BioThreat / BioDefense Program Director, LANL*
** Johar Ali, Ph.D., Cancer Genomics Team Leader, OICR*
** Patrick Chain, Bioinformatics/Metagenomics Team Leader, LANL*
** Michael FitzGerald, Microbial Special Projects Manager, Broad Institute*
** Bob Fulton, M.S., Director of Project Development & Management, WashU*
** Darren Grafham, Lab Manager, Children's Hospital, Sheffield, UK*
** Alla Lapidus, Ph.D., Associate Director, Algorithmic Biology Lab, SPbSU, Russia*
** Donna Muzny, M.Sc., Director of Operations, BCM*
** Nadia Fedorova, Genome Finishing and Analysis Team Leader, JCVI*

| 05/29/2013 - Wednesday | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| 7:30 - 8:30am | Breakfast | x | La Fonda Breakfast Buffet | Sponsored by NEB |
| 8:30 - 8:45 | Intro | x | Welcome Intro from Los Alamos National Laboratory | TBD |
| x | Session Chair | x | Session Chairs | Chair - Johar Ali<br>Chair - Donna Muzny |
| 8:45 - 9:30 | Keynote | FF0093 | Genomic Futurism | Richard Gibbs |
| 9:30 – 9:50 | Speaker 1 | FF0064 | Elucidating the Effects of the Deepwater Horizon Oil Spill on the Atlantic Oyster Using Global Transcriptome Analysis and GeneSifter | Todd Smith |
| 9:50 - 10:10 | Speaker 2 | FF0081 | New Tools for Comprehensive Exome Analyses Improves Clinical Utility of Exome Sequencing | Christian Buhay |
| 10:10 – 10:30 | Break | x | Beverages and Snacks Provided | Sponsored by DNAnexus |
| 10:30 – 10:50 | Speaker 3 | FF0137b | Roche 454 sequencing --- (Roche) | Jim Knight |
| 10:50 – 11:10 | Speaker 4 | FF0232 | Illumina Next Generation Sequencing Sample Preparation Improvements -- (illumina) | Haley Fiske |
| 11:10 – 11:30 | Speaker 5 | FF0229 | Sequencing for All, Enabled by the Ion PGM™ and Ion Proton™ Systems --- (LifeTech) | Jason Affourtit |
| 11:30 – 11:50 | Speaker 6 | FF0206 | Emerging Applications and Roadmap for the PacBio RS II --- (PacBio) | Stephen Turner |
| 11:50 - 12:40 | Panel Discussion | x | Next Generation Sequencing Technology Panel Discussion | Chair - Bob Fulton<br>Chair - Patrick Chain |
| 12:40 – 2:00pm | Lunch | x | Coronado Lunch Buffet | Sponsored by PacBio |
| x | Session Chair | x | Session Chairs | Chair - Tina Graves<br>Chair - Bob Fulton |
| 2:00 – 2:20 | Speaker 7 | FF0158 | Bioinformatic Applications of PacBio Long Reads to Genomic Sequencing and Finishing | Adam English |
| 2:20 – 2:40 | Speaker 8 | FF0039 | Optical Mapping Aids Contig Localization and Order: Lessons Learned from Applying it to the Unusual Region of Chromosome 4q in Zebrafish | Jonathan Wood |
| 2:40 – 3:00 | Speaker 9 | FF0157 | A High-Throughput Pipeline for Improving Assemblies with Optical Maps | Elliott Drábek |
| 3:00 – 3:20 | Speaker 10 | FF0124 | Self-Validating Technology-Agnostic Genome Assembly | Bud Mishra |
| 3:20 – 3:40 | Speaker 11 | FF0099 | De Novo Mapping with Single-Molecule Detection In Solid-State Detectors | John Oliver |
| 3:40 – 4:00 | Break | x | Beverages and Snacks Provided | Sponsored by BioNano |
| 4:00 - 6:00pm | Tech Time Talks (15 min each) | FF0212 | Developing 400-base Sequencing for the Ion PGM™ | Daniel Mazur |
| | | FF0031 | Irys: De Novo Assembly and Structural Variation Detection in Complex Genomes Using Extremely Long Single-Molecule Imaging | Harper VanSteenhouse |
| | | FF0044 | Genome Mapping in Nanochannel Arrays for Sequence Assembly and Structural Variation Analysis | Ernest Lam |
| | | FF0074 | 1000 Cancer Gene Panel for Clinical Next Generation Sequencing | Yilin Zhang |
| | | FF0101a | Meeting the Challenges of High-throughput Library Construction for Illumina Sequencing from Low-input and FFPE Samples: Engineered Enzymes and Optimized, Automated Protocols | Maryke Appel |
| | | FF0104 | Direct Selection of Microbiome DNA from Host DNA | Bradley Langhorst |
| | | FF0216 | qRNA-Seq™ - High Precision Gene Expression Analysis Using Molecular Indexing | Masoud Toloue |
| | | FF0156 | Getting to Q60 with Pure PacBio(r) Long Reads | David Alexander |
| 6:00 – 7:30pm | Posters - Even #s<br>Meet & Greet Party | EVEN #s | Poster Session with Meet & Greet Party (Sponsored by Roche)<br>Food & Drinks | Sponsored by Roche<br>6:00pm - 9:00pm |
| 7:30 - 9:00pm | Posters - Odd #s<br>Meet & Greet Party | ODD #s | Poster Session with Meet & Greet Party (Sponsored by Roche)<br>Food & Drinks | Sponsored by Roche<br>6:00pm - 9:00pm |
| 9:00 - bedtime | on your own | x | Night on your own - enjoy | x |

| *05/30/2013 - Thursday* | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| **7:30 - 8:30am** | **Breakfast** | **x** | **Santa Fe Breakfast Buffet** | **Sponsored by NEB** |
| **8:30 - 8:45** | Intro | **x** | Welcome Back | **TBD** |
| **x** | Session Chair | **x** | Session Chairs | Chair - Mike Fitzgerald<br>Chair - Nadia Fedorova |
| **8:45 - 9:30** | **Keynote** | **FF0121** | **Challenges and Opportunities in Strain-level Comparative Genomics** | **Mark Adams** |
| **9:30 - 9:50** | Speaker 1 | **FF0210** | Genome of the Pathogen Porphyromonas gingivalis Recovered from a Biofilm in a Hospital Sink using a New Platform and the Single-cell Assembler SPAdes | **Glenn Tesler** |
| **9:50 - 10:10** | Speaker 2 | **FF0139** | Characterization of the Culex Mosquito Virome in California by Metagenomic Sequencing | **Stanley Langevin** |
| **10:10 - 10:30** | Speaker 3 | **FF0205** | Food Security: the 100K Pathogen Genome Project | **Bart Weimer** |
| **10:30 - 10:50** | **Break** | **x** | **Beverages and Snacks Provided** | **Sponsored by OpGen** |
| **10:50- 11:10** | Speaker 4 | **FF0152** | Genomics Capability Development and Collaborative Research with Global Engagement | **Helen Cui / Tracy Erkkila** |
| **11:10 - 11:30** | Speaker 5 | **FF0010** | Establishing Regional NextGen Whole Genome Sequencing | **Adam Kotorashvili** |
| **11:30 - 11:50** | Speaker 6 | **FF0194** | Biobanking and Metagenomics Platforms for Pathogen Discovery | **George Michuki** |
| **11:50 - 1:10pm** | **Lunch** | **x** | **New Mexican Lunch Buffet** | **Sponsored by Perkin Elmer** |
| **x** | Session Chair | **x** | Session Chairs | Chair - Donna Muzny<br>Chair - Johar Ali |
| **1:10 - 1:30** | Speaker 7 | **FF0071** | The $1M Metagenomics Algorithm Challenge | **Edward Wack** |
| **1:30 - 1:50** | Speaker 8 | **FF0237** | Analytical Process for Interactive Analysis of Deep Sequencing Data on a Laptop | **Ben McMahon** |
| **1:50 - 2:10** | Speaker 9 | **FF0160** | Metagenomic Applications for Diagnostics and Etiologic Agent Discovery | **Joe Petrosino** |
| **2:10 - 2:30** | Speaker 10 | **FF0238** | Clade-specific Genomic Signatures as a Method for Accurate Profiling of Metagenomic Datasets | **Patrick Chain** |
| **2:30 - 2:50** | Speaker 11 | **FF0117** | Metagenomic and Metatranscriptomic Analyses of the Complex Community of a Tropical Wastewater Treatment System | **Stephan Schuster** |
| **2:50 - 3:10** | **Break** | **x** | **Beverages and Snacks Provided** | **Sponsored by CLCbio** |
| **3:10 - 5:00pm** | **Tech Time Talks (15 min each)** | **FF0051b** | CLC bio Products and Supported Applications | **Marta Matvienko** |
| | | **FF0075** | The CLC Microbial Genome Finishing Module | **Martin Simonsen** |
| | | **FF0030** | Latest Advances in Bioinformatics Computing | **George Vacek** |
| | | **FF0149** | Implementing Fast Sequence Analysis Tools Using a Cray XMT2 | **Sterling Thomas** |
| | | **FF0221** | Genomics Applications in the Cloud with the DNAnexus Platform | **Andrey Kislyuk** |
| | | **FF0078** | Sequence Consensus Algorithms & Hierarchical Genome Assembly Process for Effective De Novo Assembly with SMRT® Sequencing | **Aaron Klammer** |
| | | **FF0126** | Haplotype Assembly Refinement & Improvement | **Christine Olsen** |
| **5:30 - 8:00pm** | **Happy Hour** | **x** | **Happy Hour at Cowgirl Cafe - Sponsored by LifeTech<br>Map Will be Provided** | **Sponsored by LifeTech** |
| **8:00 - bedtime** | **on your own** | **x** | **Dinner and Night on Your Own - Enjoy!!!** | **x** |

| 05/31/2013 - Friday | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| 7:30 - 8:30am | Breakfast | x | Harvey House Breakfast Buffet | Sponsored by NEB |
| 8:30 - 8:45 | Intro | x | Welcome Back | Chris Detter |
| x | Session Chair | x | Session Chairs | Chair - Patrick Chain<br>Chair - Bob Fulton |
| 8:45 - 9:30 | Keynote | FF0085 | Keep Calm and Carry On as the Human Reference Assembly Updates | Deanna Church |
| 9:30 – 9:50 | Speaker 1 | FF0054 | HAVANA Manual Annotation: the Cartography of a Genome | Mike Kay |
| 9:50 – 10:10 | Speaker 2 | FF0133 | Genomic Analysis of Susceptibility to Autoimmunity | Ward Wakeland |
| 10:10 – 10:30 | Speaker 3 | FF0019 | Cross-platform NGS Method for the Identification STR Loci | Daniel Bornman |
| 10:30 – 10:50 | Speaker 4 | FF0134 | Tomorrow's Genome: Complete Bacterial Genomes in <24 h for Outbreak Response | Ken Dewar |
| 10:50 – 11:10 | Break | x | Beverages and Snacks Provided | Sponsored by AATI |
| 11:10 – 11:30 | Speaker 5 | FF0029 | Consed and BamScape for Next-Gen Sequencing | David Gordon |
| 11:30 – 11:50 | Speaker 6 | FF0137a | Assembling Human Genomes | Jim Knight |
| 11:50 – 12:10 | Speaker 7 | FF0144 | Reducing Assembly Complexity of Microbial Genomes with Single-molecule Sequencing | Adam Phillippy |
| 12:10 - 1:10pm | Lunch | x | Santa Fe Deli Lunch Buffet | Sponsored by illumina |
| x | Session Chair | x | Session Chairs | Chair - Mike Fitzgerald<br>Chair - Darren Grafham |
| 1:10 - 1:30 | Speaker 8 | FF0168 | Reference Assisted Assembly With ALLPATHSLG | Sante Gnerre |
| 1:30 - 1:50 | Speaker 9 | FF0080a | SPAdes: Assembling Microbes in the Cloud | Anton Korobeynikov |
| 1:50 - 2:10 | Speaker 10 | FF0146 | Inexpensive High Quality Genome Assemblies from a Single PCR-free Illumina Library | Ted Sharpe |
| 2:10 - 2:30 | Speaker 11 | FF0097a | Fat-Free Bioinformatics: Successful Microbial Genomics in a Lean Contract Research Environment | Jonathan Jacobs |
| 2:30 - 2:50pm | Closing Discussions | x | Closing Discussions for General Meeting - discuss next year's meeting | Chair - Chris Detter |


# Block the Noise and Focus on Your Research

## xGen™ Blocking Oligos for Target Capture

xGen™ Blocking Oligos enhance the performance of your capture experiments by preventing cross-hybridization between adapter sequences or hybridization of adapters to capture probes. By binding to the adapters, these oligos prevent non-specific hybridization to off-target sequences, increasing the number of "reads on target".

- Perform more enrichment reactions with 10 nmole or 25 nmole yields.
- All blocking oligos are compatible with the Nimblegen SeqCap SR and LR protocols.
- Customize blocking oligos for any platform and for any level of multiplexing.

*For more information go to www.idtdna.com/xgen*

IDT®
INTEGRATED DNA TECHNOLOGIES

THE CUSTOM BIOLOGY COMPANY

WWW.IDTDNA.COM

100%

| 05/29/2013 - Wednesday | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| 7:30 - 8:30am | Breakfast | x | La Fonda Breakfast Buffet | Sponsored by NEB |
| 8:30 - 8:45 | Intro | x | Welcome Intro from Los Alamos National Laboratory | TBD |
| x | Session Chair | x | Session Chairs | Chair - Johar Ali<br>Chair - Donna Muzny |
| 8:45 - 9:30 | Keynote | FF0093 | Genomic Futurism | Richard Gibbs |
| 9:30 – 9:50 | Speaker 1 | FF0064 | Elucidating the Effects of the Deepwater Horizon Oil Spill on the Atlantic Oyster Using Global Transcriptome Analysis and GeneSifter | Todd Smith |
| 9:50 - 10:10 | Speaker 2 | FF0081 | New Tools for Comprehensive Exome Analyses Improves Clinical Utility of Exome Sequencing | Christian Buhay |
| 10:10 – 10:30 | Break | x | Beverages and Snacks Provided | Sponsored by DNAnexus |
| 10:30 – 10:50 | Speaker 3 | FF0137b | Roche 454 sequencing --- (Roche) | Jim Knight |
| 10:50 – 11:10 | Speaker 4 | FF0232 | Illumina Next Generation Sequencing Sample Preparation Improvements -- (illumina) | Haley Fiske |
| 11:10 – 11:30 | Speaker 5 | FF0229 | Sequencing for All, Enabled by the Ion PGM™ and Ion Proton™ Systems --- (LifeTech) | Jason Affourtit |
| 11:30 – 11:50 | Speaker 6 | FF0206 | Emerging Applications and Roadmap for the PacBio RS II --- (PacBio) | Stephen Turner |
| 11:50 - 12:40 | Panel Discussion | x | Next Generation Sequencing Technology Panel Discussion | Chair - Bob Fulton<br>Chair - Patrick Chain |
| 12:40 – 2:00pm | Lunch | x | Coronado Lunch Buffet | Sponsored by PacBio |
| x | Session Chair | x | Session Chairs | Chair - Tina Graves<br>Chair - Bob Fulton |
| 2:00 – 2:20 | Speaker 7 | FF0158 | Bioinformatic Applications of PacBio Long Reads to Genomic Sequencing and Finishing | Adam English |
| 2:20 – 2:40 | Speaker 8 | FF0039 | Optical Mapping Aids Contig Localization and Order: Lessons Learned from Applying it to the Unusual Region of Chromosome 4q in Zebrafish | Jonathan Wood |
| 2:40 – 3:00 | Speaker 9 | FF0157 | A High-Throughput Pipeline for Improving Assemblies with Optical Maps | Elliott Drábek |
| 3:00 – 3:20 | Speaker 10 | FF0124 | Self-Validating Technology-Agnostic Genome Assembly | Bud Mishra |
| 3:20 – 3:40 | Speaker 11 | FF0099 | De Novo Mapping with Single-Molecule Detection In Solid-State Detectors | John Oliver |
| 3:40 – 4:00 | Break | x | Beverages and Snacks Provided | Sponsored by BioNano |
| 4:00 - 6:00pm | Tech Time Talks (15 min each) | FF0212 | Developing 400-base Sequencing for the Ion PGM™ | Daniel Mazur |
| | | FF0031 | Irys: De Novo Assembly and Structural Variation Detection in Complex Genomes Using Extremely Long Single-Molecule Imaging | Harper VanSteenhouse |
| | | FF0044 | Genome Mapping in Nanochannel Arrays for Sequence Assembly and Structural Variation Analysis | Ernest Lam |
| | | FF0074 | 1000 Cancer Gene Panel for Clinical Next Generation Sequencing | Yilin Zhang |
| | | FF0101a | Meeting the Challenges of High-throughput Library Construction for Illumina Sequencing from Low-input and FFPE Samples: Engineered Enzymes and Optimized, Automated Protocols | Maryke Appel |
| | | FF0104 | Direct Selection of Microbiome DNA from Host DNA | Bradley Langhorst |
| | | FF0216 | qRNA-Seq™ - High Precision Gene Expression Analysis Using Molecular Indexing | Masoud Toloue |
| | | FF0156 | Getting to Q60 with Pure PacBio(r) Long Reads | David Alexander |
| 6:00 – 7:30pm | Posters - Even #s<br>Meet & Greet Party | EVEN #s | Poster Session with Meet & Greet Party (Sponsored by Roche)<br>Food & Drinks | Sponsored by Roche<br>6:00pm - 9:00pm |
| 7:30 - 9:00pm | Posters - Odd #s<br>Meet & Greet Party | ODD #s | Poster Session with Meet & Greet Party (Sponsored by Roche)<br>Food & Drinks | Sponsored by Roche<br>6:00pm - 9:00pm |
| 9:00 - bedtime | on your own | x | Night on your own - enjoy | x |

# *NOTES*

# Speaker Presentations (May 29th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0093

**Genomic Futurism**

Richard A. Gibbs, Ph.D.
Director and Wofford Cain Professor
Human Genome Sequencing Center
Department of Molecular and Human Genetics

Predicting major societal change requires out-of-the-box speculation: In genomics, we can easily predict widespread access to DNA testing based upon whole genome sequencing. This alone can dramatically influence our lives as individuals, as well as society at large. When combined with other 'DNA related' technologies, there is enormous potential for genomics to influence future realities. Just like the proliferation of cellphones, personal computers and the Internet, these may seem like science fiction when looking ahead. An important challenge is to identify those elements of change that are reasonable enough to be seeded in current activities, but able, with some imagination, to be scaled and adapted to new influential capabilities.

# *NOTES*

FF0064

**Elucidating the Effects of the Deepwater Horizon Oil Spill on the Atlantic Oyster Using Global Transcriptome Analysis and GeneSifter**

Natalia G. Reyero (1), Nalini Raghavachari (2), Kurt Showmaker (1), Poching Liu (2), Nadereh Jafari (3), Natalie Barker (4), Kristine L. Willett (5), Jone Corrales (5), Heather K. Patterson (6), Ruth H. Carmichael (6), Don Baldwin (7), N. Eric Olson (8), Hugh Arnold (8), and Todd M. Smith (8).

1. M*ississippi State University, 2. NIH, 3. Northwestern University, 4. USACE, 5. University of Mississippi, 6. DISL, 7. Pathonomics LLC, 8. Geospiza, a PerkinElmer Company.*

The Deep Water Horizon oil spill resulted in the release of over 200 million gallons of crude oil into the waters of the Gulf of Mexico. Additionally, two million gallons of chemical were used to emulsify and disperse oil plumes posing further risks to the environment. Biota such as the commercially important Atlantic oyster *Crassostrea virginica* were inevitably exposed to spill-related contaminants. The potential effects of oiled water and sediments on oysters range from non-detectable to reduced settlement to impaired immune function, acute intoxication, and death due to bioaccumulation of contaminants. As sedentary organisms, oysters are even more susceptible to the negative effects of oil contamination. In order to understand the mechanisms of oil and spill-related compound induced toxicity, we sequenced the RNA of oyster samples from before and after the spill. A challenge with analyzing the resulting data is that an annotated C. *virginica* genome is not available. As the related Pacific oyster genome was recently released, we reasoned that its genome could be used to annotate RNA-seq data from *C.* virginica. However, individual reads from *C. virginica* mapped poorly to the Pacific oyster genome limiting the utility of this approach. To improve annotation, we performed a *de novo* transcriptome assembly to create long contigs that provided 66-70% alignment rate and thus identify 9,469 homologous transcripts between the Atlantic and the Pacific oyster. This approach can be further generalized to identify previously unannotateed but transcriptionally active regions of any genome. And, when used in conjunction with whole genome sequencing mutations in a larger repertoire of expressed regions of a genome can be identified.

FF0081

**New Tools for Comprehensive Exome Analyses Improves Clinical Utility of Exome Sequencing**

Christian Buhay, Qiaoyan Wang, Shruthi Ambreth, Mark Wang, Yi Han, Huyen Dinh, Harsha Doddapaneni, Yaping Yang, Alicia Hawes, Victor Zhang, Lee-Jun Wong, Matthew Bainbridge, Eric Boerwinkle, Jeffrey Reid, Donna M. Muzny, Richard A. Gibbs

Baylor College of Medicine, Houston, TX 77030.

Comprehensive exome analyses are hallmarks of clinical samples sequenced in the Whole Genome Laboratory (WGL). To that end, we continually develop tools to fine-tune analyses, increase probability of diagnosis, and improve clinical utility. Further improvements of complete clinical exomes occur in three phases: addition of mitochondrial genome sequencing, automatic detection of inadequately covered exons, and development of methods to rescue clinically significant low coverage regions.

To better characterize coverage across the HGSC whole exome design (VCRome2.1), targeted gene regions across 34 clinical exome samples were scrutinized. We specifically focused on the performance of relevant diagnostic genes across the Genetest and COSMIC lists. Results show that >90% of genes in the list have a base-depth coverage of 20X or better in our clinical samples. Leveraging the coverage tools developed in the process, we have implemented the Exome Coverage and Identification (ExCID) Report in our research pipeline. This workflow assesses exome coverage, annotates target regions with gene and exon coverage depth, as well as reports inadequately (<20X) covered exonic regions for every exome coming through the production pipeline. Shortly after sample sequencing and mapping, investigators can now quickly and efficiently assess performance across genes of interest. Results of the pipeline are leading to development of methods and reagents to rescue low-coverage regions in future exomes, as well as provide meta-data for other future applications.

Mitochondrial variant discovery has been missing in many exome capture designs. Starting in October 2012, the WGL has included mitochondrial sequencing as a value-added component to exome sequencing. Long-range PCR, sequencing, and analysis strategies were developed at BCM-HGSC for both the Ion Torrent and Illumina platforms. Six controls (one normal and five patient samples) were amplified and sequenced at the HGSC and WGL to a depth of 5000X or better. Such deep sequencing insures that we detect low frequency (10%) heteroplasmy and large duplications or deletions with granularity for diagnostic calls. Mutation detection was completed and all variants were compared against known samples. In every sample, regardless of sequencing platform, all known variants were validated.

FF0137b

**Roche 450 Sequencing**

Jim Knight

Roche

FF0232

**Illumina Next Generation Sequencing Sample Preparation Improvements**

Haley Fiske

Illumina

Illumina has recently launched and improved numerous sample prep kits. These kits increase the utility of the HiSeq and MiSeq DNA sequencing systems. Sample prep technologies for long reads and targeted RNA will be discussed.

FF0229

**Sequencing for All, Enabled by the Ion PGM™ and Ion Proton™ Systems**

Jason Affourtit

Director of Product Management, Library Products

Ion Torrent, part of Life Technologies

Abstract – Ion Torrent has invented the first device—a new semiconductor chip—capable of directly translating chemical signals into digital information. The Ion Personal Genome Machine™ Sequencer, launched in December of 2010, delivered 1000X scalability improvements in its first year of commercial availability. The PGM now can deliver over 2 GB of data per run using the 318 chip with 400 bp read lengths. Ion Torrent released the Ion Proton™ Sequencer in late 2012. The P1 chip routinely generates 12 GB of across its 165 million microwells with 200 bp read lengths. Both sequencers generate data for a wide variety of applications include: gene panel sequencing, exome analysis, transcript analysis (include whole transcriptome, small RNA, and targeted RNA sequencing), copy number analysis, 16S analysis, and de novo assembly. An update on these platforms, including current performance and how it enables routine genome sequencing applications, will be presented.

FF0206

**Emerging Applications and Roadmap for the PacBio RS II**

Stephen Turner

PacBio

With the recent release of the PacBio RSII (and the associated upgrade), its throughput and readlength realize several new applications that are impractical with second-generation sequencing technologies. This talk is a survey of the new applications emerging both from the customer base as well as from Pacific Biosciences' research and development efforts. In addition to the applications of genome assembly and reference sequence improvement that will be discussed separately at this conference, we'll discuss advances in full-length transcript sequencing. We'll show progress in both sample preparation and read length resulting in an increase in the proportion of reads from a SMRT® Cell that represent full-length transcript sequences. Targeted sequencing and phasing in fields such as cancer genomics and HLA typing have both benefitted from the increased read length, as well as the development of algorithms to exploit these long reads. We will conclude with an overview of the expected performance increases for the rest of the year.

# *NOTES*

# *NOTES*

# Lunch

**12:40 – 2:00pm**

## Sponsored by

# *NOTES*

FF0158

**Bioinformatic Applications of PacBio Long Reads to Genomic Sequencing and Finishing**

Adam English, Will Salerno, Alicia Hawes, Yi Han, Mark Wang, Donna Muzny, Kim Worley, Stephen Richards, Jeff Reid, Richard Gibbs

Human Genome Sequencing Center - Baylor College of Medicine - Houston, TX

The unprecedented length of PacBio reads (3 kb mean, 6 kb N50) allows for accurate mapping despite a high—but stochastic—error rate (~85% base-accuracy). PBJelly[1] leverages these data to span intra-scaffold gaps and perform a guided local assembly to fill said gaps. Here we present extensions to the PBJelly software that further exploit the unambiguous mapping of PacBio reads. Specifically, we address three persistent problems in genomic sequencing and finishing: inter-scaffold gap filling, misassembly detection, and structural variation identification. First, we fill inter-scaffold gaps by incorporating into the assembly reads that align past the ends of scaffolds. Second, we show how misassemblies can be detected via the alignment of PacBio reads. Finally, we posit that structural variations can be considered misassemblies, thus allowing us to identify structural variants that may go undetected by short-read methods.

As evidence, we apply these novel methods to a cohort of simulated and real data derived from bacterial-to-mammalian-sized genomes. With 24x mapped coverage of PacBio long reads, we close 69% and improve 12% of all gaps in *D. pseudoobscura*, increasing the contig N50 from 53 kb to 224 kb. Using 40x of PacBio reads from *E. coli*, we created several draft de-novo assemblies using various PacBio assembly techniques, scanned them for misassemblies, and filled inter-scaffold gaps. These improved assemblies have a scaffold N50 between 25 kb to 70 kb. Additionally, we randomly selected a number of human structural variants with exact breakpoints from NCBI's dbVar and simulated PacBio reads to determine the feasibility of classifying a broad range of structural variation event sizes and types (e.g., insertion, deletion, and translocation).

1. English AC, Richards S, Han Y, Wang M, Vee V, et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7(11): e47768. doi:10.1371/journal.pone.0047768

FF0039

**Optical Mapping Aids Contig Localization and Order: Lessons Learned from Applying it to the Unusual Region of Chromosome 4q in Zebrafish.**

Jonathan Wood, Kerstin Howe and Matthew Dunn

Wellcome Trust Sanger Institute, Cambridge, UK

There have been many efforts to produce an accurate map for the long arm of chromosome 4 in zebrafish, from radiation hybrid and meiotic maps to physical BAC clone fingerprint maps. These have produced poor results for this particularly complex heterochromatic region that - even in the current assembly Zv9 - is still highly fragmented with little evidence for the current contig order.
In order to improve the representation of chromosome 4, we applied OpGen's novel Genome Builder process. It utilises local single molecule assemblies from optical mapping to join sequence contigs together creating large sequence scaffolds.

The resultant local optical map assemblies span several megabase in size, providing previously absent long-range information. We used this to place sequence contigs, leading to notable inter and intra-chromosomal rearrangements in the current assembly which further aids the analysis and understanding of this region.

FF0157

**A High-Throughput Pipeline for Improving Assemblies with Optical Maps**

Elliott F. Drábek, edrabek@som.umaryland.edu

Arthur L. Delcher
Xian Fan
Anup A. Mahurkar
Luke J. Tallon
Mark Eppinger
Herve Tettelin
Emmanuel Mongodin
Claire Fraser

Institute for Genome Sciences
University of Maryland School of Medicine
Baltimore, MD

We describe a high-throughput pipeline for improving genome assemblies of bacteria and small eukaryotes using optical map data. The pipeline requires minimal user intervention and produces .agp files indicating how contigs map to optical maps, creating a single scaffold covering the entire bacterial chromosome. The pipeline incorporates genome assembly scaffold information to place contigs which otherwise might be too small to place uniquely on the optical map. It can also use aligned short-read data to extend contigs and/or close gaps using the IMAGE tool, and to correct errors in contig consensus sequences by detecting and correcting differences.

In cases where a discrepancy exists between the optical map and the assembly, such as a chimeric contig or scaffold, the pipeline flags the discrepancy for manual examination and correction. A particularly novel feature of the pipeline is the ability to combine results from multiple assemblies, choosing portions from contigs from them to create a single assembly with the most consistent and complete coverage of the optical map.

We report results for our pipeline on assemblies for a diverse set of bacterial genomes, including multiple strains of *Vibrio cholerae*, *E. coli*, and *Staphylococcus aureus*. On these test strains, the pipeline is able to span each entire genome with a single scaffold while, on average, covering 89% of the optical map and leaving 34 gaps of average length 7.9 Kb. This illustrates that optical maps can be a cost-effective tool for creating nearly finished-quality bacterial genome assemblies.

This project has been funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900007C.

FF0124

**Self-Validating Technology-Agnostic Genome Assembly**

Chang Peng, Harsh Patel, Ed Burzycki, Giuseppe Narzisi, Vadim Sapiro, Andi Witzel, Jason Reed, and Bud Mishra

MRTechnology, OpGen, New York University, Virginia ComΩmonwealth University, Google, & Cold Spring Harbor Lab

In the recent past, our DNA-sequencing capability has improved exponentially with advancements in speed, throughput, read length, and cost. However, they face many critical obstacles: (1) To be useful, these technologies must rely on auxiliary information that requires specialized sample preparation (pooling, dilution, bar-coding, etc.) or other low-resolution long- range information (mate pairs, strobed sequencing, optical maps, dilution mapping, etc.). (2) As their data throughput has increased, there is an unmet need for this data to be rapidly analyzed, compressed and stored – severely limited by available bandwidth and storage space. (3) In the absence of data standards (either for short- or long-range genomic data), the analysis step and the subsequent pipeline must be developed in a technology-agnostic manner. (4) Finally, the end results, often a sequence assembly or resequencing, need to be validated as adequately accurate for the intended scientific purposes. Consequently, the assembly pipeline, and the set of algorithms it comprises, must be smart, flexible, efficient, and scalable. Here we describe a sequencing pipeline, created collaboratively by data from MRTechnology and OpGen and consisting of TotalRecaller-SUTTA-FRCValidator-WholeGenomeMaps. TotalRecaller improves the base-calling quality using a prior, obtained from a reference genome and/or bootstrapped from the contigs as generated by earlier reads; SUTTA improves the assembly quality by using a branch-and-bound approach that optimizes Bayesian score functions, designed to self-validate the assembly using single-molecule restriction maps (from OpGen) and/or other layout-features. Current implementation is being tested on microbial genomes (E. coli and V. cholerae) with encouraging results.

FF0099

## *De Novo* Mapping with Single-Molecule Detection In Solid-State Detectors

John S. Oliver, Debra Dederich, Brendan Galvin, Peter Goldstein, William Heaton, Dona Hevroni, Leah Seward, Adam Snider, John Thompson

Nabsys Inc., Providence, RI

Next-generation sequencing systems have revolutionized the use of sequencing for answering many biological questions. However, typical read lengths from NGS technologies are too short for unambiguous assembly. Assembly of the resulting data typically generates short contigs and incomplete assemblies.

An alternative to standard finishing methods is to create *de novo* assembled physical maps that can be used as a scaffold for assembly of short read data. We have mapped DNA by attaching probes to specific regions and then using fully solid-state nanodetectors for localizing the positions of those probes. DNA entering the nanodetector causes a drop in the current with the time of the current drop dependent on the length of the DNA. The probes attached to the DNA are sufficiently large that they cause an additional alteration in the current and the timing of the current changes can be used to determine the length and relative positions of the probes. Because each DNA molecule is measured independently and can be hundreds of kilobases long, it is straightforward to assemble a complete genome for microbial species of interest. We will present results on the *de novo* mapping of small genomes with a high-throughput and scalable detector.

FF0212

**Developing 400-base Sequencing for the Ion PGM™**

Daniel Mazur, Guobin Luo, Xinzhan Peng, Anelia Kraltcheva, Tommie Lincecum, Eileen Tozer, Kristen Aguinaldo, Geoffrey Lowman, Mindy Landes, Brian Strohecker, Theo Nikiforov, Peter Vander Horn

Ion Torrent Division, Life Technologies, 5781 Van Allen Way, Carlsbad, California 92008 (Daniel.Mazur@lifetech.com)

The utility of long sequencing reads for enhanced genomic annotation and assembly has been a key differentiator for successful next generation sequencing applications. Through optimization of both the sequencing and amplification biochemistries, we have significantly improved Ion Torrent's semiconductor chip-based sequencing system. Along with improvements in engineering and software, these innovations result in robust high quality 400-base sequencing reads on the Ion Torrent PGM platform. Using this improved 400 bp biochemistry, we have produced >2G aligned Q20 bases with even coverage on the *E. coli* genome and significantly improved coverage on the more challenging *Rhodo sphaeroides* genome.

Compared to the previous PGM sequencing protocols, the Ion Torrent 400bp Template/Sequencing kits give nearly double the read length and throughput. Such improvements in the PGM system enable a broader range of applications, such as enhanced de novo genome assemblies, Human Leukocyte Antigen (HLA) sequencing, bacterial identification, and meta-genomic analysis.

*For Research Use Only. Not for use in diagnostic procedures.

FF0031

**Irys: De Novo Assembly and Structural Variation Detection in Complex Genomes Using Extremely Long Single-Molecule Imaging**

H VanSteenhouse, A Hastie, M Requa, M Austin, F Trintchouk, M Saghbini, X Yang, H Cao

BioNano Genomics
9640 Towne Centre Drive, Suite 100
San Diego, CA 92121

De novo genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in most genomes. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses.

The Irys platform from BioNano Genomics overcomes the limitations of short fragment technologies to provide unprecedented insights into whole-genome biology. Irys is a single-molecule genome analysis system based on NanoChannel Array technology that linearizes extremely long DNA molecules for observation. This high-throughput platform automates massively parallel imaging of individual molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness and long-range contiguity critical for unambiguous de novo assembly of complex genomes. High-resolution genome maps assembled de novo retain the original context and architecture of the genome, making them extremely useful for sequence assembly scaffolding and structural variation detection applications.

Specifically, these genome maps provide dense genome-wide anchor points for ordering and orienting sequencing contigs or scaffolds to greatly increase completion and accuracy of de novo assemblies. Structural variants and repeats are measured directly within long “reads” for comprehensive analysis of what has been dubbed “the inaccessible genome”. Following an introduction to the platform and underlying single-molecule technology, several examples of real-world application will demonstrate its use in large complex genomes

FF0044

**Genome Mapping in Nanochannel Arrays for Sequence Assembly and Structural Variation Analysis**

E. Lam1, A. Hastie1, M. Requa1, H. Dai1, M. Austin1, F. Trintchouk1, M. Saghbini1, T. Anantharaman1, H. VanSteenhouse1, S. Rombauts2, V. Zhurov3, K. Haden1, M. Grbic3, T. Dickinson1, X. Yang1, E. Holmlin1, H. Cao1

1. BioNano Genomics, San Diego, California, US
2. Vlaams Instituut voor Biotechnologie, Flanders, Belgium
3. University of Western Ontario, London, Ontario, Canada

Genome mapping involves direct analysis of extremely long single DNA molecules and represents a platform complementary to short-read sequencing technologies. Long-range information is preserved, enabling haplotype and structural variation detection. In the context of sequence assembly, the long molecules can bridge sequence scaffolds and span across repeat regions otherwise difficult to assemble using short reads, improving the fidelity of an assembly.

By incorporating genome mapping data, we significantly improved the draft assembly of the spider mite (*T. urticae*), a common pest and model organism for arthropod biology. We created superscaffolds of draft sequence scaffolds anchored to consensus genome maps. The scaffold N50 was improved by more than two-fold. We corrected misassembled sequence scaffolds and uncovered haplotypes with strong single-molecule support. We used the genome map data to confirm the copy number of repeat-containing silk genes in the spider mite.

We also applied genome mapping analysis to one Caucasian and one Asian genome. We constructed consensus genome maps that cover the majority of the genome. We detected large structural variants hundreds of kilobases in size and haplotype differences around the major histocompatibility complex (MHC) region. We also covered regions annotated as gaps in hg19 and estimated the gap sizes.

FF0074

**1000 Cancer Gene Panel for Clinical Next Generation Sequencing**

Jer-Kang Chen1, Lihyun Sun1, Dora Dias-Santagata2,3, I-Ching Wang1, Teresa Lin1, Darrell Borger3, Leif Ellisen3, A. John Iafrate2,3, Long Phi Le2,3*, Yilin Zhang1*

1. Elim Biopharmaceuticals, Inc., 25495 Whitesell St., Hayward, CA 94545

2. Pathology, Massachusetts General Hospital, Boston, MA

3. Cancer Center, Massachusetts General Hospital, Boston, MA

Cancer genotyping in the clinical setting demands high coverage to achieve the analytical sensitivity for detecting low frequency variants in heterogeneous specimens. At the present time, this is only achievable in practice with sub-exome targeted gene panels. We have developed a broad, 1000 cancer gene panel for next generation sequencing based on enrichment by liquid phase hybridization and targeting all coding exons and intron-exon junctions of the selected genes. The design of the panel contains 3 tiers of genes related to cancer: (1) known oncogenes and tumor suppressors that are clinically actionable, (2) additional oncogenes and tumor suppressors for which clinical utility have not been demonstrated, and (3) other tyrosine kinase, phosphatase, and vascular genes.

The assay has been developed comprehensively as a library construction and capture kit (RightOn Cancer Sequencing Kit, Elim Biopharm) which will accommodate small quantities of formalin-fixed paraffin-embedded clinical DNA samples. Performance of the assay shows minimum 100X coverage for more than 98% of the target bases at a sequencing throughput of 5 gigabase per sample. Such high on target coverage allows robust detection for single nucleotide variants, insertions/deletions, and copy number changes. Data will be presented for a variety of clinical specimens. Sequencing in conjunction with cancer drug screening through primary cell lines from cancer patients in mice will also be presented. To our knowledge, this is the largest cancer gene panel developed so far with a focus on high coverage for sensitive detection of low frequency variants. In contrast to current commercially available exome capture kits which show incomplete coverage for many genes, this kitted assay offers a nearly complete and high coverage cancer sequencing option for clinical application and discovery.

FF0101a

**Meeting the Challenges of High-throughput Library Construction for Illumina Sequencing from Low-input and FFPE Samples: Engineered Enzymes and Optimized, Automated Protocols**

Maryke Appel, Kapa Biosystems Inc., Woburn, MA

Innovation in Next-Generation Sequencing (NGS) has been focused on core sequencing technologies, with the optimization of library preparation playing a secondary role. We previously reported on the advantages of engineered DNA polymerases for the quantification and quality assessment of input DNA (particularly FFPE samples), low-bias library amplification (of both normal and bisulfite-converted DNA), and accurate quantification of NGS libraries prior to pooling for capture or sequencing. To extend the benefits of low-bias amplification, particularly to low-input applications and challenging samples such as FFPE, we have developed, optimized and automated a core DNA library construction protocol for Illumina sequencing that incorporates the "with-bead" strategy conceptualized by The Broad Institute of MIT and Harvard. Together with the use of ultra-pure, high quality reagents for library construction, our highly optimized "with-bead" protocol results in significantly higher recoveries of adapterligated molecules. This allows for robust library construction from lower amounts of input DNA and FFPE samples and/or fewer cycles of amplification, thereby further reducing the risk of PCR-induced bias, error and other artefacts that can affect library and sequence quality. The protocol was designed for a seamless transition from low-throughput, manual library construction to high-throughput production pipelines. Automated methods for the three major liquid handling platforms used in high-throughput NGS library construction have been developed and validated with different sample types, and for different sequencing applications. We will be presenting data from several collaborators to illustrate the benefits of our engineered enzymes and optimized library construction protocol for targeted sequencing, low-input applications and FFPE samples.

FF0104

**Direct Selection of Microbiome DNA from Host DNA**

Erbay Yigit George R Feehery, Bradley W Langhorst, Lynne M Apone, Pingfang Liu, Daniela B Munafó, Christine J Sumner, Joanna Bybee, Laurie M Mazzola, Fiona J Stewart, Theodore B Davis, Eileen T Dimalanta, Sriharsa Pradhan

New England Biolabs, Inc., 240 County Rd, Ipswich, MA 01938

Nucleic-acid based techniques such as hybridization, PCR, qPCR and next generation sequencing offer a rapid and highly sensitive option for direct metagenomic detection from specimens when compared with culture-based techniques. Currently, 16S rRNA gene sequencing is the method of choice for microbiome studies. However, this approach lacks sensitivity to detect a rare member of the microbial community with divergent target sequences, and is not adequate to detect virulence factors in individual strains. Aside from these inherent limitations of amplification and identification of biological samples, the non-microbial host genome itself may interfere with the detection and diagnosis of pathogens due to the higher percentage of host genomic DNA relative to the target microbiome. Therefore, analyses of a microbiome directly from host samples by next generation sequencing or PCR are inefficient, difficult and time consuming. To address this problem, we have developed a unique method for separating large pieces of host DNA from microbial DNA using a methyl-CpG binding domain fused to the Fc portion of a human antibody heavy chain (MBD2-Fc). This MBD2-Fc protein is then bound to a Protein A magnetic bead, and used to separate methylated host DNA from unmethylated microbial DNA. As a demonstration of the efficacy of this methodology, DNA samples from various sources were enriched and sequenced on different next generation sequencing platforms. Sequencing data showed that non-microbial host DNA reads decreased 50-fold, whereas microbiome DNA reads increased 9-fold, corresponding to ~90-95% microbiome DNA in the enriched fraction. Importantly, microbiome diversity after the enrichment remained intact. This simple methodology can be used to analyze entire microbiomes in a cost-effective manner utilizing established next generation sequencing platforms, as well as newer single molecule sequencing technologies.

FF0216

**qRNA-Seq™ - High Precision Gene Expression Analysis Using Molecular Indexing**

Masoud Toloue

Bioo Scientific, 3913 Todd Lane Suite 312, Austin, TX 78744, USA

Most modern methods for NGS library prep require the use of enzyme processing, such as DNA polymerase reactions, which can introduce errors in the form of incorrect sequence and misrepresented copy number. Conventional RNA sequencing library construction involves the ligation of a population of cDNA molecules with adapters prior to amplification and sequencing. An inherent weakness of conventional RNA-Seq analysis is that cDNA fragments that amplify more efficiently will unavoidably result in a higher number of reads than cDNAs that do not amplify as well during the library construction PCR step. Therefore, when multiple reads mapping to the same transcript are encountered, it is not possible to determine whether sequenced reads originate from the same or different cDNA molecules. With Molecular Indexed™ libraries, each molecule is tagged with a molecular index randomly chosen from ~10,000 combinations so that any two identical molecules become distinguishable (with odds of 10,000/1), and can be independently evaluated in later data analysis. Analysis using molecular indexing information provides an absolute, digital measurement of gene expression levels, irrespective of common amplification distortions observed in many RNA-Seq experiments. This type of indexing requires no additional steps in RNA-Seq workflow and increases the precision of downstream analysis.

At low sequencing depths, analysis using the NEXTflex qRNA-Seq Kit is identical to conventional analysis and generates equivalent RPKM values in all applications. As sequencing depth increases, individual molecular resolution also increases. In quantitative RNA-Seq experiments, the molecular indices distinguish re-sampling of the same molecule from sampling of a different molecule. At high sequencing depths, each molecule can be distinguished and the entire library can be analyzed to provide absolute numbers of each molecule. Resolving individual clones of molecules is critical for increasing sequencing accuracy or when identifying mutations in complex sample types.

FF0156

**Getting to Q60 with Pure PacBio(r) Long Reads**

David H. Alexander
Senior Algorithms Engineer
Pacific Biosciences

Patrick Marks
Staff Engineer
Pacific Biosciences

The error structure of the long reads generated using PacBio's single-molecule real-time (SMRT®) sequencing technology is fundamentally different from that of short-read technologies. In PacBio reads, sequencing indels are the most common error type, while substitutions are rare. Bioinformatics tools for consensus and variant calling that do not account for this different nature of PacBio reads will thus deliver suboptimal results. In this talk, we describe how the accurate modeling of the error modes in SMRT sequencing has led to the development of Quiver, an algorithm for highly-accurate consensus and variant calling that lies at the core of PacBio's latest resequencing and assembly polishing workflows.

Applications to microbial genome assemblies have shown that genome consensus sequences generated by Quiver are as accurate as those generated using Sanger-sequencing approaches, at a fraction of the cost. Comparison to alternative variant callers has proven that a model that correctly accounts for the error structure of the sequence affords greater sensitivity in variant detection.

# *Notes*

# *Meet and Greet Party*

600pm – 900pm, May 29th

Sponsored by Roche Diagnostics

Enjoy!!!

# *Poster Presentations (May 29th)*
## *Even #'s 6:00-7:30pm, Odd #'s 7:30-9:00pm*

FF0011

**JCVI Viral Finishing Pipeline: Improvements and Challenges**

Nadia Fedorova, Danny Katzel, Timothy B. Stockwell, Peter Edworthy, Rebecca Halpin, and David E. Wentworth

The J Craig Venter Institute, Rockville, MD, 20850 U.S.A.

JCVI viral projects are supported by the NIAID Genomic Sequencing Center for Infectious Disease (GSCID). The viral sequencing and finishing pipeline at JCVI combines next generation sequencing (NGS) technologies with automated data processing. This has enabled the completion of over 12,300 viral genomes since 2005, including over 11,000 Influenza genomes.

Our NGS pipeline uses SISPA, Nextera, and/or Ion Torrent library construction methods coupled with Illumina and/or Ion Torrent PGM sequencing. Our automated assembly pipeline employs CLCbio command-line tools and JCVI cas2consed, and is capable of assembling input data from any sequencing platform, both with and without references. JCVI autoTasker provides automated quality control and validation functionalities. Fully automated assembly pipeline is integrated with JCVI's LIMS and JIRA Workflow Management system.

The availability of small-scale next generation sequencing instruments has opened new opportunities for targeted finishing. Converting routine closure sequencing from Sanger to Ion Torrent PGM will speed turnaround time while decreasing the cost of finishing. Our goal is to reserve Sanger sequencing with improved primer design for limited challenging applications.

While our pipelines are geared towards high-throughput processing, they also have the ability to sequence, assemble, and finish novel viruses, repetitive or recombined genomes, and samples containing multiple viruses. This flexibility comes from optional manual reference selection and reference editing steps. Our pipeline has been used to complete many novel adenovirus genomes and avian influenza co-infections. As an example, we are presenting an overview of the sequencing, assembly, and finishing of a novel bat coronavirus genome, including the challenges presented by working with novel RNA virus directly from a fecal sample.

In summary, the JCVI automated viral pipeline, fully integrated with lab management and tracking software, follows the progress of viral samples from acquisition through to NCBI submission. This allows us to process a large volume of samples with limited manual interaction, while giving us flexibility to work on challenging and novel genomes.

FF0013

**Annotation and Comparative Analysis of *Lactuca sativa* and its Wild Progenitor *Lactuca serriola* using CLC Genomics Workbench**

Alexander Kozik, Dean Lavelle, Sebastian Reyes-Chin-Wo, Lutz Froenicke, Richard Michelmore

UC Davis Genome Center, Davis, CA

The genome of cultivated lettuce *Lactuca sativa* cv. Salinas has been sequenced in collaboration with the BGI and a consortium of ten breeding companies (http://lgr.genomecenter.ucdavis.edu/). To understand the evolutionary events underlying lettuce domestication, we have initiated a comparative analysis of *L. sativa* with its closest wild relative and likely progenitor *L. serriola*. We sequenced genomic libraries of *L. serriola* and compared coverage of mapped reads across the whole genome of *L. sativa*. Genomic reads from a large RIL population of a *L. sativa* x *L. serriola* cross were used to confirm the differences in genome representation in these two *Lactuca* species. In addition, RNA-Seq libraries from both species were used to assist in the analysis. Using CLC Genomics Workbench we analyzed coverage for single copy genes and multi-gene families. CLC Genome Finishing Module was used to identify and annotate highly conserved and diverged (fast evolving?) genome regions in the *L. sativa* assembly. CLC Transcript Discovery plugin assisted with finding and annotation of transcribed regions that were missed in the first round of prediction of translated genes. Multiple types of annotation were integrated, visualized and analyzed simultaneously using CLC Track Tools. Comparative studies on the complete *L. sativa* and *L. serriola* genome assemblies is ongoing. Ultimately, this will lead to understanding of the genome architecture of transcribed sequences, the organization and evolution of gene families, as well as the role of repetitive and transposable elements in *Lactuca* species.

FF0023

**Whole Genome Shotgun Sequencing and *Denovo* Assembly of the Atheriniform Fish *Odontesthes bonariensis* (Pejerrey)**

Campanella, Daniela[12] ; Miller, Jason[1]; Somoza, Gustavo[3]; Fernandino, Juan[3]; Ortí, Guillermo[2]; Caler, Elisabet[1]

[1.] J. Craig Venter Institute, [2.] The George Washington University, [3.] IIB-INTECH (CONICET-UNSAM, Argentina)

*Odontesthes bonariensis*, commonly known as *pejerrey,* inhabits freshwater environments of the South American Pampas. The pejerrey is a popular sport fishery species, with a long history of domestic and international introductions due to the high quality and market value of its flesh. Several efforts for its aquaculture have been made for over a century in South America and Japan.

A first draft of the pejerrey genome was obtained using whole genome shotgun sequencing of 3 Illumina libraries of 200, 300 (paired-end) and 3 K bp (mate-paired) insert size. Assembly of the 1 billion resulting reads was attempted using SOAPdenovo and AllPaths. The AllPaths assembly resulted in larger number of bases incorporated in a smaller number of scaffolds with higher N50. The SOAP assembly resulted in 147,183 scaffolds, 32.5 X coverage and N50=32kb (contig N50=2,553 in 189,087 contigs). The AllPaths assembly resulted in 31,274 scaffolds, 40 X coverage, and N50=61kb (contig N50=13kb in 148,185 contigs). The estimated genome size is 800 Mb.

Preliminary assembly analyses showed a low repeat content, and long stretches of di-nucleotide repeats, characteristic of vertebrate microsatellites. *Swimmer,* a LINE family of retrotransposons, originally described in medaka and desert pupfish, was also identified. Preliminary alignments with medaka chromosomes showed high sequence identity and synteny: 21,562 medaka genes have orthologous genes in the pejerrey genome, revealing a great level of conservation despite 80M years of divergence between Atheriniformes (pejerrey) and Beloniformes (medaka).

FF0024

**Leveraging the CLC Bio Platform for NGS Analysis – Automate It Your Way**

Cecilie Boysen, Marta Matvienko, et al. CLC bio, Cambridge, MA, US

CLC bio is known for its ease of use, graphical displays, and fast and accurate de novo and mapping algorithms for NGS reads used in the data analysis of different applications. These range from whole human genome, exome, or transcriptome analysis, deep amplicon sequencing used in cancer, mitochondrial or viral sequencing, to microbial or large genome assemblies. As these applications have moved from the research stage and the 'one sample at a time' scenario, into a production line environment, more and more users are now using the CLC bio platform to automate their preferred workflows for high-throughput analysis and sharing of data.

Here we present the CLC bio software platform, including the CLC Genomics Server and Workbench client and different ways users interact with the system to automate and customize their analyses. This varies from the bio-informaticians, who integrate their own algorithms and use command-line tools to generate automated workflows, to the biologists, who make use of the new workflow build and management tool to build their customized. These workflows can be locked down as often required for standardized analyses used in high-throughput clinical, forensics and biosurveillance labs.

Examples of automatic workflows from raw input data to final results will be presented for applications such as mutation detection and structural variation in cancer to viral and bacterial analysis solutions.

FF0038

**Assessing the Population Structure of Fungal Pathogens using Whole-genome Sequence Data**

N. Hicks, C. Roe, J. Gillece, J. Schupp, E.M. Driebe, P. Keim, D. Engelthaler
TGen North, Flagstaff, AZ

Next-generation sequence data has become a valuable tool in infectious disease epidemiology, allowing for the evolutionary history of pathogenic microbes to be incorporated into traditional epidemiological analysis. For a growing number of bacteria and viruses, genetic variation data, frequently measured using SNP loci, have allowed disease outbreaks and individual cases to be linked to previously sequenced source strains through phylogenetic analyses. Fungal pathogen whole-genome SNP data are becoming increasingly available, promising to extend molecular epidemiology analyses to fungal outbreaks; however, fungi can propagate both clonally like bacteria and through sexual recombination like other eukaryotes. Thus, a full understanding of fungal evolutionary history can only be described by both the clonal relationships within strain types and the population structure that encompasses strain types. Population structure is commonly assessed using distance-based methods, such as principle coordinates analysis (PCA), which do not incorporate an explicit model of evolutionary relationship, and model-based methods such as Structure. While Structure has been effective for relatively small datasets, next-generation sequencing of large fungal collections has produced very high-density SNP data, which cannot be effectively processed by Structure and violates model assumptions of unlinked markers. Here we present population genomic analyses of two fungal pathogens of interest, *Coccidioides spp.* and *Cryptococcus gattii*, using high-density SNP data from whole genome sequences. Using phylogenetic models, PCA, and the recently released model-based analysis tool, fineStructure, we explore the ability of each method to provide population structure information. fineStructure promises to efficiently capture more information from high-density SNP data by explicitly using the tight linkage between markers that are physically close on chromosomes to increase the sensitivity of population structure detection. In addition, we use inferred structure as a framework for exploring large genomic differences that population subdivision can create by allowing for local adaptation and genetic drift.

FF0040

**Facing the Challenge of High-Throughput Finishing**

Guy Griffiths, on behalf of the Genome Reference Consortium

Wellcome Trust Sanger Institute, Cambridge, England

The Sanger-sequencing and finishing of clone sequences is a well-established process, successfully applied for more than a decade. However, with the advent of new sequencing technologies, offering remarkably increased throughput at decreased cost, we face a great opportunity to adapt and improve this system.

This challenge was taken up by the Sanger Institute by firstly changing from Sanger-sequencing of individual clones to sequencing indexed pools of up to 96 clones per lane on Illumina HiSeqs or MiSeqs, developing gap5 to handle both Sanger and short-read-based projects and developing a pipeline to assemble the data (as shown at SFAF meeting in 2012).

Secondly, the finishing process was adapted by using increased automation and additional QC, significantly reducing the time needed for manual intervention. The assessment, ordering and orientating of sequence for a BAC clone insert (HTG phase II) now takes approximately 30 min, allowing a single finisher to elevate more than 80 clones a month into a stage that can be included in reference assemblies golden paths, e.g. those provided by the Genome Reference Consortium (GRC, genomereference.org) for the Human, Mouse and Zebrafish genomes. The quick turnaround of phase II sequences also allows the rapid identification of clones with features of interest and subsequent fast elevation of these clones into phase III (gold standard sequence).

The newly developed procedures have greatly benefitted the GRC reference assemblies by enabling the sequence curators to quickly react to error reports and sequencing requests from the research communities and by increasing the amount of data available for inclusion into assembly updates within a given time frame and budget.

FF0046

## *C. Elegans* Database of Evolutionarily Young Gene Duplicates: A Simple Way to Visualize Duplicated Gene Structures and Gene Clusters

Lijing Bu[1], Vaishali Katju[1]
1. Department of Biology, University of New Mexico

Gene duplication is a universal event that acts as a driving force in the evolution of genomes, from smaller, simpler ones to larger, complex ones. In addition to the addition of an extra copy by *complete* gene duplication that has potential to enable the novel evolution of one copy under relaxed selective constraints, *partial* and *chimeric* gene duplications, which were neglected in early studies, also occur at high rates and have the potential to develop new functions. Although gene duplication databases were established recently to automatically identify and record gene duplications from sequenced genomes, a database delineating the structural relationship between duplicated gene pairs has yet to be developed. In this study we propose to establish a database that allows users to visualize and compare the gene structure and flanking regions for gene duplicates within the *Caenorhabditis elegans* genome. Protein sequences of canonical transcripts for each protein coding gene were clustered based on sequence similarity using CD-HIT. Small gene families that have less than five members were selected. ORFs and 50 kb flanking sequences of genes paralogs were aligned using LASTZ program to find regions of high similarity. These alignments were then filtered and loaded into a preinstalled local synteny database of the GBrowse_syn program. The General Feature Format (GFF3) for *C. elegans* was also uploaded into the database for annotation visualization. This *C. elegans* database for evolutionarily young gene duplicates will provide direct visual observation of the duplication boundaries, exon structural similarity, and duplicated gene clusters within this genome.

FF0050

## *De novo* Assembly of Genomes Using Large Insert Pacific Biosciences SMRT Sequencing Data

A. Clum[1], A. Copeland[1], L. Hickey[2], M. Ashby[2], D. Alexander[2], P. Marks[2], J. Chin[2], L. Pennacchio[1]

1Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California, 94598, USA; 2Pacific Biosciences, 1380 Willow Rd., Menlo Park, CA 94025

Widespread adoption of short-read technologies initially created challenges for de novo assembly, but new assembly algorithms designed to work with these data rapidly reduced the performance gap, and make possible high quality assemblies using only short-read data. We have previously reported on a hybrid assembly approach combining deep coverage 2nd generation short-read data with Pacific Biosciences SMRT reads and showed that many gaps in 2nd generation data could be closed with data from 3rd generation single-molecule sequencing. These hybrid assembly approaches typically rely on short-read for error correction of the longer SMRT reads. Recently it has been shown that short CCS reads can be used for error correction instead of high coverage short-reads. In this we will describe assembling microbial and fungal genomes solely from reads of an 8-10kb library sequenced on the Pacific Biosciences platform. Using data from finished genomes allows us to assess the contiguity, accuracy, and correctness of annotations based on the resulting contigs. We will show that it is possible to accurately reconstruct genomes in a small number of pieces using this approach.

FF0051a

**CLC bio Tools for Microbial Genome Assembly and Finishing**

Marta Matvienko[1], Martin Simonsen[2], Poul Liboriussen[2], Peder Roed Lindholm Nielsen[2], Jesper Jakobsen[2], Steffen Mikkelsen[2], Henrik Sandmann[2], Søren Mønsted[2], Jannick Dyrløv Bendtsen[2]

[1]CLC bio, USA, [2]CLC bio, DENMARK

*De novo* genome assembly and genome finishing is becoming increasingly important with the massive amounts of data being generated by next generation sequencing. The NGS methods still leave room for Sanger sequencing data in finishing projects for validation and closure. The hybrid data sources can complement each other to generate high quality assemblies.

Here we demonstrate how an assembly of a microbial genome can be optimized using the CLC Microbial Genome Finishing Module. The module is a collection of tools for identifying, visualizing and solving problems in genome assemblies from both NGS and Sanger reads. The publicly available sequence data for *Pseudomonas aeruginosa* MPAO1 was used to optimize the parameters for genome assembly and finishing. The contigs were assembled using CLC bio *de novo* assembler. To evaluate the assembly, and to design the reagents for genome closure, we used Microbial Genome Finishing Module.

We optimized the word and bubble sizes for this dataset to obtain the initial assembly with the best possible quality for the given datatset. This step allowed us to reduce the number of contigs from 48 to 23. The contigs were analyzed using the *Analyze contigs* tool in the Finishing Module. This identified and annotated the problematic regions that needed further attention. Those were the regions with low, high, single-stranded, unstable coverage and regions with unaligned read ends. After manually inspecting and editing mappings in these areas, we aligned the resulting contigs to *Pseudomonas aeruginosa* PAO1 genome assembly. The alignment of contigs to this close reference identified overlapping contigs, which were joined using the corresponding tool. To close the gaps in the genome, we designed primers for Sanger sequencing using the automated primer design tool. The primers were designed for the ends of all contigs and for the regions with low coverage.

**Conclusion** Optimizing *de novo* assembly parameters allowed us to significantly improve the assembly and reduce the number of contigs. Microbial Genome Finishing Module produced a higher-quality genome assembly than *de novo* assembler alone.

FF0052

**Whole Genome Mapping Enables Improved and Validated New Technology Sequencing Projects: A Brief Overview of the Technology and Associated Mapping Projects at Sanger Over the Last Year.**

Matt Dunn[1], Michelle Smith[1], Richard Rance[1]

[1]Wellcome Trust Sanger Institute

Whole Genome Mapping is a process which allows the creation of a genome or chromosome sized high resolution restriction map of an organism, from very small quantities of high molecular weight DNA. Utilising the Argus platform from OpGen, DNA is captured in parallel arrays of single DNA molecules using a microfluidic device, stained, digested and visualised under fluorescent microscopy. Post digestion the individual fragments within molecules of DNA are then accurately measured and the captured data assembled together according to restriction cutsite patterns, thus creating a de novo restriction enzyme map. Sequence data can be subsequently digested *in silico* and aligned to the map, allowing order, orientation and gap sizing information to be inferred, together with independent sequence assembly validation.

Utilising this approach we have generated whole genome restriction maps of numerous and diverse organisms, from bacteria through to large vertebrates, enabling significant enhancements in the associated sequence assemblies and the construction of more accurate and complete genome architecture. This outlined process is fully illustrated through the review of data from several genome projects, demonstrating that a combined sequence and mapping approach can empower high quality assembly of new technology sequencing data.

FF0053

**Process for Complete Viral Genome Sequencing**

Michael G. FitzGerald, Malboeuf C, Levin JZ, Newman RM, Zody M, McCowan C, Ireland A, Fan L, Qu J, Yang X, Charlebois P, Ryan E, Poon T, Murphy C, Birren B.

Viral illness exacts a massive global health burden. Changes to local environments, urban expansion, increased global travel and other factors are leading to expanded range for some of these pathogens. Emerging viral threats like SARS accentuate the need for robust surveillance and tracking systems. The Broad Institute, through our NIAID funded Genome Sequencing Center for Infectious Disease program, seeks to apply genome science to viral disease. Advancing sequencing and analysis technologies allow us to explore the diversity and epidemiology of these viruses in unprecedented detail. Our RNA virus sequencing programs, Dengue (DENV), West Nile (WNV) and Hepatitis C (HCV), have traditionally generated whole genome assemblies through targeted PCR amplicon-based approaches. This whole-genome assembly approach offers a richer data set compared with the more typical targeted sequencing approaches that rely on small, predefined diagnostic regions. We have recently explored the use of sequence-independent amplification methods (*Malboeuf et al*, NAR), notably the NuGEN Ovation RNA-Seq system. We will detail both our amplicon-based and sequence-independent viral amplification systems, note advantages and limitations of both systems and provide data on the sequence performance for each.

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services under Contract No.: HHSN 272200900018C

FF0061

**Generation of High-quality Draft Assemblies with a Single Sequencing Library**

Sarah Young, Scott Steelman, Riza Daza, Marc Chevrette, Robert Lintner, Sante Gnerre, Aaron Berlin, Bruce Walker, Chad Nusbaum and Robert Nicol

Broad Institute of MIT and Harvard, Cambridge MA

Generating highly accurate and contiguous *de novo* genome assemblies from short-read sequencing technologies requires paired read data from multiple libraries providing short- and long-range linking information. While sequencing costs have continued to decrease over time, library construction costs remain comparatively high. Typically, fragment and jumping libraries are prepared separately despite many similarities in the two methodologies, increasing not only the amount of sample DNA required, but also the associated labor and reagent costs.

We sought to develop a method that allows for the simultaneous preparation of Illumina short-insert read pairs along with longer distance mate-pair libraries in a single round of library construction which will make efficient usage of sample DNA and decrease labor and reagent cost. In standard jumping library preparation, small insert fragments arise naturally when circular segments are sheared, however, they are normally discarded during the process. In this method, we add a molecular tag to the ends of large (~5 kb) DNA fragments as they are circularized allowing mate-pair junction sequences to be distinguished from fragment sequences.

By identifying the location of the molecular tag at the circularization junction, we are able to computationally separate out the jumping pairs from the local fragment pairs. These processed files can be fed into our standard assembly pipelines, and we have produced high-quality assemblies using these data for a fraction of the cost of our standard assemblies.

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.:HHSN272200900018C

FF0065

**A Fast Solution to NGS Library Prep with Low Nanogram DNA and RNA Input**

Pingfang Liu, Daniela B Munafo, Gregory JS Lohman, Eric Cantor, Bradley W Langhorst, Erbay Yigit, Lynne M Apone, Christine Sumner, Thomas C Evans Jr., Nicole Nichols, Fiona J Stewart, Eileen T Dimalanta, and Theodore B Davis

As use of NGS is becoming more widespread, more challenging samples are being used, including samples available only in small amounts. At the same time, demand is increasing for faster protocols that work reliably and do not compromise the quality of the libraries produced.

To overcome these challenges for DNA, we have developed a fast library preparation method using novel reagents, including a new DNA polymerase that has been optimized to minimize GC bias. This method enables library construction from amounts as low as 5 ng, and is suitable for both intact and fragmented DNA.

Standard “next-generation” RNA-sequencing approaches erase RNA strand information. The polarity of the transcript is important for correct annotation of novel genes, identification of antisense transcripts with potential regulatory roles, and for correct determination of gene expression levels in the presence of antisense transcripts. To address this need we have developed a novel streamlined, low input method for Directional RNA-Sequencing that highly retains strand orientation information while maintaining even coverage of transcript expression. This highly streamlined method is based on second strand labeling and excision after adaptor ligation; allowing differential tagging of the first strand cDNA ends.

FF0070

**Genetic Characterization of an Atypical *Vibrio cholerae* Isolate from Iraq**

Maryann Turnsek[1], Lee Katz[1], Lawrence Hon[2], Ellen E. Paxinos[2], Yan Guo[2], Susana Wang[2], Mike Frace[1], Muthana M. Al-Shemri[3], John D. Klena[4], and **Cheryl L. Tarr[1]**

[1] Centers for Disease Control and Prevention, Atlanta, GA, USA; [2]Pacific Biosciences, Menlo Park, CA ; [3]Central Public Health Laboratory, Baghdad, Iraq; [4]US Naval Medical Research Unit No. 3, Cairo, Egypt.

*Vibrio cholerae* is a Gram-negative bacterium that is the causative agent of cholera. Historically, pandemic *V. cholerae* O1 strains that encode cholera toxin (CT) have been categorized into two biotypes, classical (CLA) and El Tor (ET). Prototypical CLA and ET strains have specific *rstR* sequences associated with their respective CT phages (e.g., $rstR^{ET}$ with $CTXf^{ET}$). In recent years, isolates with characteristics of both CLA and ET biotypes have been found ('atypical' or 'hybrid' El Tor). These isolates differ from prototypical CLA and ET strains in the organization, copy number, and/or sequence of phage genes. We used PCR to survey a collection of recent *V. cholerae* isolates for phage typing in the CTXf region and discovered 2011EL-1144H, an isolate recovered from a case in Baghdad, Iraq in 2008. The isolate yielded amplicons for $rstR^{Cla}$, $rstR^{ET}$, and $rstR^{Calcutta}$; the latter was initially described from an O139 strain from Calcutta, India. We further characterized the CTXf region(s) of the isolate by PCR and Sanger sequencing, and determined the whole genome sequence (WGS). Illumina and 454 sequencing was performed at CDC with reads of size 70bp and 318+/132bp respectively, and Single Molecule Real Time (SMRT®) Sequencing was performed by Pacific Biosciences. Sanger sequencing of *rstR* amplicons confirmed the classification of the three allele types. The *ctxAB* sequence matched CLA, which is consistent with currently circulating atypical El Tor strains. The phylogenetic analysis based on SNPs from WGS clustered the isolate with other 7$^{th}$ pandemic isolates, including additional Iraq outbreak isolates from 2007 and 2008. CTX Phage-associated genes were found in both chromosomal insertion sites, with CLA, ET, Calcutta *rstR* alleles on chromosome 1 and CLA sequences associated with chromosome 2. The application of PCR assays that exploit polymorphism in the CTXf region are useful for identifying atypical isolates that may warrant further characterization. The long read SMRT sequencing was critical for resolving the arrangement of repeat units in the phage arrays.

FF0072

**Genome Map Assembly from Nanochannel Array Data for Structural Variation Detection in the Human Genome and Finishing in *Tribolium***

W. Andrews1, E. Lam1, A. Hastie1, H. Dai1, M. Coleman2, M. Austin1, F. Trintchouk1, M. Saghbini1, T. Anantharaman1, H. VanSteenhouse1, K. Haden1, T. Dickinson1, S. Brown2, X. Yang1, E. Holmlin1, H. Cao1

1. BioNano Genomics, San Diego, CA 92121, USA
2. Division of Biology, Kansas State University, Manhattan, KS 66506, USA

We present the use of a newly available technology utilizing NanoChannel Arrays to analyze complex genomic architecture and functional regions by visualization of 100 kilobase and longer strands of intact genomic DNA. A successful *de novo* assembly of the human genome is presented and utilized in structural variation analysis. Structural variants are detected as low-scoring regions of the assembly flanked by high-scoring alignments. Examples of structural ariants are presented from the KIR region of chromosome 19 and the IGH region of chromosome 14. These loci are important in the immune system function and both are known to be highly variable. The extremely long DNA molecules provide unique opportunities to study these complex structural variants which are difficult to analyze using sequencing alone.

The high assembly quality achievable with genome maps makes them useful for finishing where gaps between sequence contigs exist. We present an analysis of the *Tribolium castaneum* genome, and demonstrate the sizing of several gaps, as well as ordering and placing contigs with previously unknown locations. This organism's assembly has over 400 scaffolds and 7000 contigs and is complicated by a large fraction of repetitive heterochromatin sequence.

The Irys platform from BioNano Genomics overcomes the limitations of short fragment technologies to provide unprecedented insights into whole-genome biology. Irys is a single-molecule genome analysis system based on NanoChannel Array technology that linearizes extremely long DNA molecules for observation. This high-throughput platform automates massively parallel imaging of individual molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness and long-range contiguity critical for unambiguous *de novo* assembly of complex genomes. High-resolution genome maps assembled *de novo* retain the original context and architecture of the genome, making them useful for sequence assembly scaffolding and structural variation detection applications.

FF0080b

**HammerIT: Homopolymer-space Hamming Clustering for IonTorrent Read Error Correction**

Anton Korobeynikov1;2, Artem Tarasov1

1 Faculty of Mathematics and Mechanics, Saint Petersburg State University, Saint Petersburg, Russia
2 Algorithmic Biology Laboratory, Saint Petersburg Academic University, Saint Petersburg, Russia

Error correction of sequenced reads remains a di_cult task, especially for data obtained using IonTorrent technology due to its higher error rate. The task is even more challenging in single-cell sequencing projects with extremely non-uniform coverage.

The existing error correction tools assume that the most sequencing errors in the data are mismatches and thus perform poorly on IonTorrent data with its prevailing errors due to homopolymer indels. We introduce several novel algorithms based on homopolymer-space Hamming graph clustering in our new error correction tool HammerIT which is speci_cally tuned for IonTorrent sequencing errors.

We benchmark HammerIT on k-mer counts, read error rate and actual assembly results. Detailed analysis will be presented.

FF0088

## TAMARA: Transcriptome Analyses Based on MAssive Sequencing of RNAs

Cyrille Longin1, Marion Weiman2, David Roche1, Rachel Torchet1, David Vallenet1, Claudine Médigue1, Stéphane Cruveiller1
1 CEA/DSV/IG/Genoscope & CNRS UMR8030 - Laboratoire d'Analyses Bioinformatiques pour la Génomique et le
Métabolisme. Evry, France. 2 CNRS - Centre de Génétique Moléculaire, Département Dynamique et Stabilité des Génomes. Gif-sur-Yvette, France

The recent progress in high-throughput sequencing technologies with cost decrease led to an easier and widespread use of RNA-Seq analysis. Due to more and more user requests, the MicroScope platform dedicated to microbial genome annotation and comparative analysis [1], now includes a tool that focuses on transcriptomic data analyses based on deep sequencing of mRNAs: TAMARA. The current version of our analysis pipeline embeds the following third-party software: (i) SSAHA2 [2], BWA [3] and Bowtie [4] for the read mapping step, (ii) Samtools [5] GenomicFeatures (RBioconductor) [6] for the file manipulation and (iii) DESeq (R-Bioconductor) [7] for differential expression analyses. Moreover, the pipeline is able to handle paired-end reads, metatranscriptomic data and can manage several projects in parallel.

The generated RNA-Seq data are integrated with all other MicroScope data to extend the analysis possibilities. The gateway between the main annotations and the RNA-Seq databases enables the use of other MicroScope tools and thus provides insights into metabolic pathways or orthologous genes searching for instance. Additional links to the Integrative Genomics Viewer [8] and the Multiexperiment Viewer software [9] are provided for statistical and classification analyses, such as clustering or gene-set enrichment. Most of the RNA-Seq results can be browsed online or download locally. Though TAMARA is already functional and available, its development is still ongoing. Techniques specifically developed in order to sequence the 5' region of RNA molecules study allow the identification of potential Transcriptional Start Sites (TSSs) for experiments designed. A local peak detection algorithm is used for the TSSs prediction on these experiments taking into account the background noise. Then, it could be used in combination with transcript assembly to draw operon maps of bacterial species. Finally, in addition to differential analysis, mapping results and TSSs may contribute to significantly improve gene annotation in our MicroScope database. Currently 4 public projects can be browsed at: http://www.genoscope.cns.fr/agc/microscope/expdata/rnaseqProjects.php

**References**

[1] Vallenet, David et al. “MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data”. *Nucleic acids research* 41.Database issue (2013): D636-647.
[2] Ning, Z, A J Cox, et J C Mullikin. “SSAHA: a fast search method for large DNA databases”. *Genome research* 11.10 (2001): 1725-1729.
[3] Li, Heng, and Richard Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics (Oxford, England)* 25.14 (2009): 1754–1760.
[4] Langmead, Ben et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. *Genome biology* 10.3 (2009): R25. [5] Li, Heng et al. “The Sequence Alignment/Map format and SAMtools”. *Bioinformatics (Oxford, England)* 25.16 (2009): 2078-2079.
[6] M. Carlson, H. Pages, P. Aboyoun, S. Falcon, M. Morgan, D. Sarkar and M. Lawrence. “GenomicFeatures: Tools for making and manipulating transcript centric annotations”. R package version 1.4.3(2011)
[7] Anders, Simon, et Wolfgang Huber. ”Differential expression analysis for sequence count data”. Genome biology 11.10 (2010): R106.
[8] Thorvaldsdóttir, Helga, James T Robinson, and Jill P Mesirov. “Integrative Genomics Viewer (IGV): highperformance genomics data visualization and exploration.” *Briefings in bioinformatics* 14.2 (2013): 178–192.
[9] Saeed, Alexander I et al. “TM4 microarray software suite.” *Methods in enzymology* 411 (2006): 134–193.

FF0089

**Next Generation Sequencing in the National Health Service (NGS in the NHS)**

Darren Grafham, Head of Laboratory Services, Sheffield Diagnostic Genetics Service

Sheffield Diagnostic Genetics Service is the most integrated NHS diagnostic genetics department in the UK offering over 90 diagnostic tests covering both cytogenetic and molecular genetic disciplines to the UK and internationally. The laboratory employs over 80 staff and has a number of ongoing collaborations in research and education with partners in the public and private sectors. The Laboratory holds full CPA Accreditation, is a member of the UK Genetic Testing Network (UKGTN) and participates in external quality assessment schemes **UK NEQAS** and EMQN**. We have developed a semi automated pipeline to develop our next generation sequencing capabilities which is fully supported by our LIMS (StarLims). This will enhance patient diagnosis and care.**
**We are due to launch next generation sequencing (NGS) service panels on the MiSeq (Connective tissue disorders and glycogen storage diseases) and PGM (BRCA1/2) this spring.**
**Further development of the pipeline and the challenges of moving NGS into a clinical setting will be explored.**

FF0097b

**Towards an Improved High Quality Draft Genome of *Geobacillus stearothermophilus* Donk 1920 ATCC7953**

Jonathan Jacobs [1]*, Nicole Waybright [1], Danielle Swales [1], Brittney Knight [2], Richard Winegar [2]

* Cooresponding Author

[1] MRIGlobal, Rockville, MD
[2] MRIGlobal, Palm Bay, FL

*Geobacillus stearothermophilus* is a spore-forming, aerobic or facultatively-anaerobic, motil, chemo-organotrophic, obilgate thermophilic bacteria commonly found in warm soils, hot springs, and ocean sediment. *G. stearnothermophilus* also is commonly used as a challenge organism for confirming the proper functioning of autoclave and sterilization equipment. Recently, *G. stearnothermophilus* has shown promise as a source of novel enzymes, such as *Bst* polymerase, and other proteins for high-temperature applications in biotechnology. Dispite the potential for numerous other recombinat proteins in *G. stearnothermophilus* with application in biotechnology, no draft genome has been published or officially released publically. Here we present for the first time the high-quality draft genome of *G. stearnothermophilus* Donk 1920 ATCC7953 (*Gbs*). Whole genome sequencing was carried out using a combination of next-generation sequencing (NGS) platforms and DNA libraries, including paired-end sequencing on Illumina MiSeq, and both single-read and mate-paired sequencing on Ion Torrent PGM. *De novo* assembly of NGS data was done using CLC Genomics Workbench 6. *Gbs* consists of a genome of approximately 2.8 Mb with a 52% GC content. Our current draft assembly of Gbs consists of 89 contigs with an N50 of 87,243 bases, and ranging from 1,212 bases to 247,953 bases. Mean coverage depth on a contig – by – contig basis ranges from 44x to 1,243x, with a median coverage of 88x across all contigs. Using CLC bio Genome Finishing Tools, this draft assembly was improved to 84 contigs with the longest contig extended to 390kb. Finally, this improved draft was submitted the RAST pipeline for automated annotation and functional analysis. RAST identified and assigned putative functions to 3,126 candiate ORFs and 90 RNA genes. Future work will include additional mate-pair sequencing to create longer scaffolds of ordered contigs, and capillary sequencing to close gaps and resolve repeat regions. Submission of this improved high-quality draft genome to GenBank is pending.

FF0097c

**Genotypic and Quasispecies Analysis of Rift Valley Fever Virus Following *in vitro* and *in vivo* Passaging**

Jonathan Jacobs [1]*, Danielle Swales [1], Colby Layton [2]

* Cooresponding Author
[1] MRIGlobal, Rockville, MD
[2] MRIGlobal, Kansas City, MO

Rift Valley Fever Virus (RVFV) is a CDC Category A Select Agent. It is a highly pathogenic zoonotic arbovirus that presents a serious concern to national bio- and agricultural defense. In the event of a potential RVFV outbreak in North America, it will be critical to determine the source of the virus and distinguish between natural and laboratory raises isolates. RNA viruses, such as RVFV, are known to accumulate genetic variation due to an error prone viral RNA-dependent RNA polymerase. Accumulated changes at both the consensus and quasispecies level facilitates host adaptation and often plays important roles in viral pathogenesis, host-pathogen interaction and innate immune response, and viral escape from therapeutic challenge. These same alterations could also be potentially exploited for forensic attribution purposes following the intentional release of a biothreat agent. With this in mind, we aim to identify these molecular signatures for rapid source attribution of RVFV via next generation sequencing. In this study, we present preliminary results of Ion Torrent PGM sequencing of serial passaged RVFV obtained from tissue culture and animal model systems. RVFV was passaged sequentially and alternatively *in vitro* in *Chlorocebus sabaeus* and *Aedes albopictus* cell lines (vero and C6-36 respectively). In addition, RVFV was serially passaged *in vivo* between *Culix tarsalis* mosquitos and *Rattus norvegicus* to simulate the natural transmission cycle in the wild. Additional intraspecies passaging was also carried out in series to fully explore the genome plasticity of RVFV. In each series, up to 20 passages were conducted. Viral RNA was obtained, characterized, and subjected to next-generation RNA sequencing (RNA-Seq) to determine the consensus sequence and the degree of intrahost variation observed. The parental RVFV strain (ZH501) and every $5^{th}$ passage (where possible) of each serial passaging experiement was subjected to RNA-seq and the accumulation of both dominant and minor SNPs was assessed.

FF0097d

**Development of an Amplicon Sequencing Panel for Antibiotic Resistance Marker Detection for *Bacillus anthracis***

Jonathan Jacobs [1]*, Nicole Waybright [1], Danielle Swales [1]

* Cooresponding Author
[1] MRIGlobal, Rockville, MD

*Bacillus anthracis* is a spore-forming zoonotic pathogen that occasionally infects humans, causing cutaneous, intestinal, or pulmonary forms of anthrax. Although all three human disease forms are rare, the potential for using *B. anthracis* as a biological weapon is recognized as a serious threat. The most often recommended antibiotic for the treatment of anthrax, ciprofloxacin (CIP), remains effective as natural resistance is undocumented. However, the possibility of fluoroquinilone (FQ) resistance is a major concern escalated by research that has produced high-level CIP resistant mutants in vitro.

Most fluoroquinolone-resistant mutants have amino acid changes in quinolone resistance-determining regions (QRDRs) of the GyrA subunit of gyrase and the ParC subunit of topoisomerase IV. However, recent reports suggest additional contributing factors to FQ resistance that have yet to be identified. Next-generation sequencing has facilitated the identification of genome-wide single nucleotide polymorphisms (SNPs) and insertions-deletions (indels) in CIP resistant mutants which may reveal genetic variations responsible for FQ resistance in *B. anthracis*.

An extensive literature search identified 24 SNPs and indels in 12 distinct regions of the *B. anthracis* genome that conferred CIP resistance in vitro. This project targeted those regions by PCR and employed next-generation amplicon sequencing with the ultimate goals of (1) being able to reliably sequence, detect, and identify the antibiotic resistance regions starting with purified DNA extracts from crude environmental samples and (2) identifying the Bacillus species being sequenced.

Limit of detection studies demonstrate that targeted regions can be sequenced with sufficient coverage with only 1 genome per microliter of d-ames DNA in “dirty” extracted environmental eluate (5ul per PCR reaction). Initial studies aimed at combining the 12 PCR target region reactions into a single, highly multiplexed PCR reaction have proven successful. Future studies include further optimization of the multiplex PCR, analysis of amplicon sequencing with competing near neighbor DNA, and a final validation.

FF0101b

**An Automated, High-Throughput Library Construction Protocol with Benefits for Low-Input Applications**

Maryke Appel1, Olaf Stelling2, Olga Aminova3, Sasinya Scott4, Adriana Heguy3, Michael Berger4 and John Foskett1

*1Kapa Biosystems, Woburn, MA*
*2Alpaqua Engineering, LLC, Beverly, MA*
*3Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY*
*4Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY*

We have developed a highly optimized, automated method for high-throughput library construction using the KAPA HTP Library Preparation Kit on the Biomek FX(p) Laboratory Automation Workstation (Beckman Coulter). This method implements library construction from fragmented dsDNA to amplified library in one run, while supporting optional size selection and adapter indexing. By taking advantage of the flexibility and extendability of the Biomek software, several workflow options are available to the user in an intuitive, easy-to-use interface. For optimal results, the method employs unique automation components, designed specifically to handle small volumes of samples and precious, temperature-sensitive reagents. The performance of the automated KAPA method was compared to that of a manual method previously optimized at MSKCC. To this end, libraries were constructed for targeted Illumina sequencing from 250 ng input DNA, using either tumor cell line DNA or FFPE samples. In both cases, the automated method yielded libraries of comparable quality and complexity with two fewer cycles of library amplification than the manual method. We also investigated whether the performance of the automated method was consistent over a range of low DNA inputs. The average percentage of input DNA converted to adapter-ligated molecules, as well as the enrichment factor achieved during library amplification, was high and consistent for libraries constructed from 100 ng, 50 ng or 10 ng high quality commercial human genomic DNA. Taken together, results confirmed that the automated method and KAPA library construction reagents offer robust, high-throughput library construction over a range of DNA inputs and sample types.

FF0101c

**Co-optimization of Probes and Polymerases to Drive the Evolution of Targeted Sequencing**

Maryke Appel1, Daniel Burgess2, Michael Brockman2, John Foskett1, Dawn Green2, Bronwen Miller1, Eric van der Walt1, Jennifer Wendt2

*1Kapa Biosystems, Woburn, MA*
*2Roche NimbleGen, Madison, WI*

The evolution of targeted enrichment methods for high-throughput sequencing applications has focused optimization efforts onto a small number of persistent technical impediments to increased throughput and performance. Primarily, these include inefficiencies in library construction, automation-unfriendly workflows, and the well documented tendency of many DNA polymerases to introduce strong biases against AT- and GC-rich targets. Targeted enrichment workflows are particularly susceptible to bias due the requirement for both a pre- and post-capture amplification step. By combining enzymes specifically tailored for high performance NGS applications with innovative oligonucleotide probe designs, we developed protocols for improved library construction and targeted enrichment. The result is a workflow that retains input sample complexity, minimizes amplification biases and artifacts such as PCR duplicates and chimeric library inserts, is compatible with manual or high-throughput workflows, and delivers unparalleled sensitivity for variant discovery over a wider range of targets throughout the genome than had previously been demonstrated. We applied the new reagents and protocols to a series of difficult enrichment targets, including human exomes, and panels of genomic targets specifically designed to measure GC-dependent performance during targeted Illumina sequencing. The results presented here extend the capabilities of targeted sequence enrichment, a method that has already transformed the work of genome analysis, and will enable the future discovery of more variation, in more regions of the genome, in more samples, and in less time.

FF0101d

**Rapid, Automated Illumina Library Construction with the KAPA Library Preparation Kit on the Apollo 324™ NGS Library Prep system**

Maryke Appel1, Oanh Nguyen2, Xiaohong Fan2, Christopher Odom1, John Foskett1 and Ryan Kim2

*1Kapa Biosystems, Woburn, MA*
*2DNA Technologies and Expression Analysis Cores, UC Davis Genome Center, Davis, CA*

Kapa Biosystems has developed a core DNA library construction kit and optimized protocol which enables robust library construction for Illumina sequencing from lower amounts of input DNA, and difficult samples such as FFPE, ChIP-DNA and samples from organisms with extreme (very AT- or GC-rich) genomes. The protocol has been successfully automated on the Sciclone NGS Workstation (PerkinElmer), Biomek FX(p) Laboratory Automation Workstation (Beckman Coulter), NGS Workstation Option B (Agilent Technologies) and epMotion 5075TMX (Eppendorf). The purpose of this study was to optimize library construction using the KAPA Library Preparation Kit on the Apollo 324™ NGS Library Prep System (IntegenX), thereby offering the benefits of the KAPA reagents to lowerthroughput users with access to this platform. The system uses readily available consumables, thereby enabling users to employ library construction reagents other than those supplied with the instrument.

Optimization and validation of the KAPA Library Preparation protocol for the Apollo 324™ system entailed minor modifications to the system and reaction setup, to provide for formulation differences between the KAPA library preparation reagents and those supplied with the system. Libraries were subsequently prepared from two bacterial genomic DNA preparations (representing AT-rich and GC-rich genomes, respectively), either with the original reagents and protocol, or the protocol adapted for KAPA reagents. Libraries were prepared in duplicate from different amounts of sheared DNA (ranging between 1 ng and 300 µg), and yields of adapter-ligated libraries were quantified with the KAPA Library Quantification Kit for the Illlumina® platform. Adapter-ligated libraries prepared from lower inputs were amplified with the engineered KAPA HiFi HotStart DNA Polymerase, or Phusion® DNA Polymerase. Both amplified and unamplified samples were subjected to Illumina® sequencing. Comparative data from the libraries prepared with the KAPA *vs.* original reagents will be presented.

FF0102

**Evaluation of Hypervariable 16S and ITS Tag Sequencing on Illumina MiSeq.**

Julien Tremblay[1], Kanwar Singh[1], Alison Fern[1], Edward S Kirton[1], Feng Chen[1], Shaomei He[1], Tanja Woyke[1], Janey Lee[1], Robin A Ohm[1], Matthias Hess[123] and Susannah G Tringe[1].

[1]:DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA, 94598
[2]:Systems Biology & Applied Microbial Genomics Laboratory, Washington State University
[3]: Chemical and Biological Process Development Group, Pacific Northwest National Laboratory

In recent years, microbial community surveys extensively relied on 454 pyrosequencing technology (pyrotags). The Illumina sequencing platform HiSeq2000 has now largely surpassed 454 in terms of read quantity and quality with typical yields of up to 600 Gb of paired-end 150 bases reads in one 18 day run. Illumina recently introduced the new mid-range MiSeq sequencing platform which gives an output of more than 1 Gb of paired-end 250 base reads in a single day run. With its moderately-high throughput and support for high multiplexing (barcoding), this platform represents a suitable alternative for projects needing more conservative throughput, for instance population profiling based on prokaryotic 16S rRNA genes or eukaryotic Internal Transcribed Spacer (ITS) regions.

A workflow was therefore developed to take advantage of the Illumina MiSeq platform as a suitable tool to accurately characterize microbial communities. We surveyed microbial populations coming from various environments by targeting the 16S rRNA hypervariable regions V4, V7-V8 and V6-V8 which generated amplicons size of about 290, 307 and 460 bp respectively. These amplicons were sequenced with a MiSeq instrument from both 5' and 3' ends with a 2x250 bases sequencing configuration followed by *in silico* assembly using their shared overlapping part when applicable. Corresponding V6-V8 pyrotags data were also generated to assess validity of itag data. We also explored amplicons sequencing of the ITS2 region and examined how this marker gene can be used for fungal profiling.

Although it generates shorter reads than the 454 platform, MiSeq combines well balanced throughput with a remarkably low error rate. Our results suggest that the itags community surveys on MiSeq successfully recapture known biological results and should provide a useful tool for both prokaryotic and eukaryotic community characterization.

FF0106

**Polymerase Errors and the Accuracy of Sequencing**

Marta Orlikowska, Dominika Borek, Zbyszek Otwinowski

University of Texas, Southwestern Medical Center

Sequencing of the human genome (3 Mbp) using Sanger sequencing machines took 3 years and cost about $300 million. With next-generation sequencing (NGS) technologies, it can be achieved in weeks for 100,000 times less. Polymerase chain reaction (PCR) is the essential step in NGS library preparation procedures. It is used for target enrichment or as a fast and reliable method for preparation of indexed sequencing libraries.

In a perfect PCR reaction, all the replicated molecules would be the exact copies of the original template and the number of molecules would be doubled after each replication cycle. In practice, there are many sources of problems, e.g.: (1) the DNA polymerases used in the PCR make errors with some frequency, which may be dependent on many factors such as the local sequence environment of a DNA fragment sequence; (2) the DNA template may be damaged, which will affect the outcome of polymerase-catalyzed reaction; (3) Sequence composition and length of DNA fragments may cause preferential PCR amplification.

We performed an experiment using PCR-amplified and PCR-free libraries to investigate how different DNA polymerases behave when they encounter DNA fragments containing damaged nucleotides. We designed 48 bp oligonucleotides of known sequence, with the most common modifications. All sequencing libraries were constructed using the standard Illumina sample-preparation protocol and were tested on an Agilent Bioanalyzer chip to determine the size of the library. PCR-amplified libraries have been amplified using the Kapa HiFi (Kapa Biosystems) polymerase. Amplified and no-PCR libraries have been quantified by qPCR and sequenced using the Illumina platform.

We will present the comparison of sequencing data from PCR-amplified and PCR-free libraries to show how the KapaHiFi polymerase used in the library preparation and Bst polymerase used in the Illumina sequencing react to damaged DNA bases. This will provide an experimental base for a statistical model correcting for polymerase errors during sequencing.

FF0108

## 400 Base Pair Reads Improve Utility of Ion Torrent PGM in Microbiome Applications

Ginger A. Metcalf[1], Embriette R. Hyde[2, 3], Huyen Dinh[1], HarshaVardhan Doddapaneni[1], Eric Boerwinkle[1, 4], Joseph F. Petrosino[2, 3], Donna M. Muzny[1], and Richard A. Gibbs[1]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; [2]Department of Molecular Virology & Microbiology, Baylor College of Medicine, Houston, TX, 77030; [3]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, Texas; [4]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX

The Human Genome Sequencing Center currently uses the Ion Torrent Personal Genome Machine (PGM) for a variety of amplicon sequencing projects. However, until recently, amplicons larger than 220 bp required shearing prior to sequencing on the PGM, limiting16S rRNA gene sequencing options for microbiome studies. Previously, we evaluated 16S V4 amplicon sequencing on the PGM. Samples were derived from a study examining the impact of IgA deficiency on murine intestinal microbiota where 454 data had highlighted significant differences in the communities isolated from IgA deficient compared to wild-type mice. Amplicons from the same DNA extracts were sequenced in a multiplexed PGM run. These amplicons were sheared and size selected to meet insert size requirements. The data were analyzed using QIIME to evaluate differences in the relative abundance of the microbial taxa present in each sample. As found in the original 454 results, the PGM data revealed significant differences between the IgA-/- and WT microbial communities. However, while the trends in community shifts revealed by 454 and PGM were similar, there were differences in the data that may have been reflective of the shorter PGM reads and/or the shearing of the original amplicons prior to sequencing, both of which can impact sequence classification. Recently, the 16S experiment was repeated using the 400bp long read kit through early access collaborations with Life Technologies. These longer reads eliminated the need to shear the 291 bp V4 amplicon increasing protocol flexibility by allowing the traditional library ligation process to be replaced with a more cost effective amplicon based approach to add adapters and molecular barcodes. We found that the long read protocol yielded results virtually indistinguishable from the original 454 data when examining alpha and beta diversity between the samples. Additional samples previously sequenced on 454 are in progress to further validate the utility of the Ion Torrent PGM in microbiome applications.

FF0110

**Whole-proteome Phylogeny of Prokaryotes by Variable-length Exact-sequence Matches**

Raquel Bromberg, Zbyszek Otwinowski
University of Texas, Southwestern Medical Center
Dallas, TX

Advances in sequencing have allowed a large number of genomes to be sequenced in their entirety, with many more genomes to come in the near future. As more genomes are sequenced, many current tools for phylogeny will break down. Most alignment-based methods will not scale with the growth of information. Additionally, phylogenies constructed from whole genomes or proteomes are more robust than phylogenies done from single genes or sets of genes; for instance, they are more resistant to horizontal gene transfer. As whole genome sequences continue to accumulate, methods must be developed that are scalable and whose results can be trusted within some reasonable measure. We present a new, whole-proteome, alignment-free method for constructing prokaryotic phylogenies. Unlike many other methods including most that depend upon sequence alignments, the input set for our method can be arbitrarily large. We have run our method on ~150 archaea and ~2000 bacteria separately and compared our classification with other methods, including other whole-genome or proteome methods, the NCBI taxonomy, and the 16S rRNA reference tree, and have found the resulting trees in agreement, with our tree closer to the reference tree than trees produced by other methods.

FF0114

**Genomic and Proteomic Analysis of Five Strains of Diazotrophic, $O_2$-evolving Cyanobacteria of the Genus *Cyanothece***

Louis A. Sherman[1ξ] Uma K. Aryal[2##], Stephen J. Callister[2], Lee-Ann McCue[2], Jana Stöckel[3], Michelle Liberton[3], Benjamin H. McMahon[4], Sujata Mishra[1], Xiaohui Zhang[1], Carrie D. Nicora[2], Theress R. W. Clauss[2], Thomas E. Angel[2], Joseph Brown[2], David W. Koppenaal[2], Richard D. Smith[2], Himadri B. Pakrasi[3],

[1]Dept. Biological Sciences, Purdue University, West Lafayette, IN, 47907; [2]Pacific Northwest National Laboratory, Richland, WA 99352, USA; [3]Dept. Biology, Washington University, St. Louis, MO 63130, USA; [4]Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos Natl. Lab, Los Alamos, NM; [#]Present address: Dept. of Biochemistry, Purdue University, West Lafayette, IN 47907.

Members of the cyanobacterial genus *Cyanothece* exhibit considerable variation in physiological and biochemical characteristics. Comparative assessment of genomes across 6 *Cyanothece* strains (ATCC51142, PCC7822, PCC7424, PCC7425, PCC8801, and PCC8802), the last 5 of which were sequenced by DOE JGI, revealed many interesting differences related to protein divergence (Bandyopadhyay, Elvitigala et al. 2011). To examine whether similar differences are observed experimentally, we compared expressed proteomes of 5 strains (excluding PCC8802), and identified expression of nearly 40% of each strain's predicted proteome, with the highest number of orthologs (1818 proteins) expressed in *Cyanothece* PCC7424 and lowest number (1572 proteins) in *Cyanothece* PCC7822. Considerable differences were also evident across strains, notably in *Cyanothece* PCC7425, which showed expression of the highest number of unique proteins (682), providing direct experimental evidence of faster protein divergence in this strain. In addition to shared phycobilisome (PBS) proteins, the expression of 3 phycobiliproteins and 2 PBS-linker polypeptides was detected in *Cyanothece* PCC7424, PCC7822 and PCC8801, suggesting their possible link to phycoerythrin (PE) assembly. While homologs of the photosystem I reaction center, PsaAB were identified in all the strains, a second similar copy of PsaB (Cyan7822_4989) was identified only in PCC7822. Expression of many translated pseudogenes and hypothetical proteins underscore a need for proteomics-aided genome refinement and functional annotation to better understand protein divergence in each strain. Overall, proteomic footprints of these strains contain important biological information pertaining *Cyanothece* ecology and evolution.

The functional classification of genes identified by proteomics data was performed using Sequedex (http://sequedex.lanl.gov), which is an annotation-independent method of ascribing gene function. This enabled a broad profiling of the protein complement of the five investigated *Cyanothece* strains according to the subsystems defined by the SEED project (www.theseed.org). To provide further context and distinguish regulatory differences from genome inventory differences, 94 draft and completed cyanobacterial genomes were downloaded from NCBI in March of 2013 and profiled using the same annotation independent decomposition.

FF0123

**Conditions on "Finishable Read Sets" for Automated De Novo Genome Sequencing**

Asif Khalak[1], Ka Kit Lam[2], Greg Concepcion[1], David Tse[2]

1 Pacific Biosciences, Menlo Park
2 University of California, Berkeley

In this study, we explore the performance of *de novo* genome-assembly approaches in terms of the requirements on the amount and types of reads required for automated genome assembly and finishing. Recent theoretical work expresses minimum coverage requirements for automated *de novo* finishing with perfect reads, accounting for Lander-Waterman gap coverage and genome complexity. A simulation of the base generation process was developed to produce synthetic reads in the format of PacBio® sequencing reads, but with controlled inputs of single-pass accuracy, read-length distribution, and sample artifacts. Exploring the assembly performance of this model for different genomes defines conditions in terms of read length and coverage for "finishable read sets" over which complete, automated reconstruction is possible. The combination of read noise, genomic-repeat content, and homology are quantified in terms of their effects on the noiseless repeat spectrum, which suggests an extension of the theoretical coverage requirements to noisy reads. Assembly results from experimental data are compared the results from simulations as varied over parameters. The eventual goal of this type of analysis is to develop a more general and ultimately a unified framework to understand, design and optimize assemblers in a principled way.

FF0125

**Survey of the Microbial Diversity of Ephemeral Streams of the McMurdo Dry Valleys**

Wolf, C.R., Takacs-Vesbach, C.D., Van Horn, D.H.
University of New Mexico

The ephemeral streams of the McMurdo Dry Valleys in Antarctica are the only source of free moving water in this seemingly inhospitable environment and are epicenters of biological activity as well as essential components in cycling nutrients and solutes throughout the ecosystem. The microbial mats these streams support represent the largest biomass within the Dry Valleys and are largely composed of primary producers, bacteria and diatoms that regulate ion and nutrient concentrations. Although many studies have looked at the hydrologic activity of these seasonal streams and their biologic role in the Dry Valleys, the taxonomic composition of microbes present has not been previously studied. This study surveys the biodiversity of bacterial communities in 12 streams throughout the Dry Valleys, looking at sediment and microbial mat samples collected during the peak of the stream flow season. Using 454 pyrosequencing techniques we examined the bacterial community composition and structure in regards to stream geochemistry and activity and analyzed the taxonomic dispersal and abundance within stream microenvironments and across individual streams.

FF0127

**Bacterial Biodiversity and Function in a Cold Desert Ecosystem**

Cristina Takacs-Vesbach[1], David Van Horn[1], Heather Buelow[1], Theresa A. McHugh[2], and Egbert Schwartz[2]

[1]Department of Biology University of New Mexico Albuquerque, NM 87131
[2]Department of Biological Sciences Northern Arizona University 86011

For decades the soils of the McMurdo Dry Valleys, Antarctica were thought to be essentially aseptic. We now know that this is an ecosystem that is dominated by microorganisms, but early cultivation efforts failed to detect the apparent high diversity of the region's poorly weathered, low organic-matter soils. Initial surveys of microbial diversity using 16S rRNA gene sequencing revealed a surprising bacterial richness, including representatives from at least ten different phyla, and a high proportion of unique and rare sequences. Yet, these microbial diversity surveys were not exhaustive and little information was gained about the function of the detected microorganisms. Furthermore, given the low rates of microbial activity and decomposition rates, the question of whether this richness represents functioning vs dormant members of the community has been raised. We have conducted an exhaustive survey of the microbial richness, function, and activity of soil bacteria across gradients of moisture and salinity in the McMurdo Dry Valleys using pyrosequencing of 16S rRNA bacterial tag-encoded FLX amplicons (bTEFAP) and environmental DNA (metagenomics) combined with metatranscriptomics, stable isotope probing, and microbial activity assays. Our metagenomic and metatranscriptomic analysis represents a first step in linking community diversity and function. Comparisons of the microbial communities detected by both methods reveal a soil biodiversity that is dominated by Actinobacteria, Proteobacteris, Firmicutes, and Acidobacteria. A majority of the metagenomic and metatranscriptomic sequence was assignable to a putative function, including a large proportion of metabolic genes. Microbial function will be discussed with the goal of understanding soil ecosystem function in the McMurdo Dry Valleys and the role of bacteria in cold arid soils in general.

FF0142

**The Evolution of Microbial Species – A Look through the Genomics Lens**

Neha Varghese, Supratim Mukherjee, Natalia Ivanova, Konstantinos Mavromatis, Nikos C Kyrpides and Amrita Pati. DOE Joint Genomic Institute, Walnut Creek, CA, USA.

A genome is one time---point on the species---evolution timeline. Evolution is continuous and pervasive, a single snapshot capturing different species in different stages of evolution. In the absence of precise phenotypic traits, definition of species in microbes is based on homology of 16s rRNA genes and DNA---DNA hybridization (DDH) rates. Since evolution is manifested in specific genes as well as in large segments of genomes, homology of a single gene does not always effectively capture the divergence between two genomes. Average Nucleotide Identity (ANI) as a measure of genomic distance is explored in this work. In addition to correlating strongly with DNA---DNA hybridization rates, ANI takes into account entire genomic content and can be computed between any two sequenced genomes, agnostic of whether the genomes in question can be cultured. Finding the closest relatives of a novel organism via ANI is trivial. It is free from anomalies in the existing taxonomy and provides insight into natural groupings of organisms. While ANI---based species associations affirm existing taxonomic species definitions in the majority of cases, putative anomalies in species definitions are identified in several species, while continuums in the species space are identified in others. Disagreements between ANI---based species definitions and taxonomic species definitions are considered in further detail and putative species assignments for genomes without species definitions are proposed based on ANI---propagated associations. Finally, ANI---based genome---groups are proposed as candidates for construction of pangenomes in isolate genomes.

FF0145

**Composition of the Maize Endophytic Microbiome is Correlated with Maize Genotype**

Surya Saha[1], Alice C.L. Churchill[1], Santiago X. Mideros[1], Peter Balint-Kurti[2] and Rebecca J. Nelson[1]

[1] Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, NY, USA 14853
[2] USDA-ARS Plant Science Research Institute, Department of Plant Pathology, North Carolina State University, Raleigh, NC, USA 27695

All plants contain endophytes that have the potential to provide fitness benefits to their hosts by increasing tolerance to environmental stressors, boosting plant nutrition and growth, and providing increased resistance or tolerance to insect pests and plant pathogens. We are characterizing endophytic populations inhabiting aboveground maize tissues with the goal of associating maize genetic variation with the diversity, structure and constitution of maize-associated microbial communities. Nine maize lines, representing a diverse subset of the founders of the NAM (Nested Association Mapping) population, were field-grown and assayed for culturable endophytic bacteria and fungi. Approximately 65% of the samples contained one or more phenotypically distinct, culturable bacteria, 28% contained one or more fungi, 22% contained both bacteria and fungi, and endophytes were undetectable in 28% of the samples. Interestingly, a greater number and diversity of fungi were cultured from tropical maize lines than from temperate lines. Bacteria were isolated from all maize lines, with some lines exhibiting significantly greater microbial community diversity than others. Several phenotypically similar bacteria and fungi were isolated from multiple maize lines. Microbial cataloging of unculturable endophytes via 16S and ITS sequencing, as well as identification of novel endophytes via whole genome metagenomic sequencing, is in progress. In a parallel analysis, whole genome shotgun sequences generated for the nine maize lines were selected from the HapMap2 dataset for *in-silico* taxonomic classification of the microbial population. Identification of the bacterial microbiome is underway using FCP (naïve Bayes) and Phymm (Hidden Markov Models). Characterization of fungal endophytes is being done using a read-mapping algorithm and custom ITS databases. We are particularly interested in identifying members of the microbiome that modulate disease symptoms caused by maize leaf and ear pathogens. Hence, future studies will focus on *in vitro* and *in planta* endophyte-pathogen interactions.

FF0147

**Methods of Detecting and Characterizing Variation in *Mycobacterium tuberculosis***

Terrance Shea[1], Sarah Young[1], Sean Sykes[1], Sante Gnerre[1], Aaron Berlin[1], Sakina Saif[1], Amr Abouelleil[1], Alma Imamovic[1], Chad Nusbaum[1], Ashlee Earl[1], Bruce Walker[1],

Broad Institute of MIT and Harvard, Cambridge MA

*Mycobacterium tuberculosis* (Mtb) infects one third of the world's population and kills 1.7 million per year. It is estimated that 20% of Mtb infections are resistant to at least one drug and multi- and even totally- drug resistant strains are emerging due to the inability to effectively diagnose and treat patients. Recognizing the urgent clinical need for better diagnostics, NIAID has funded the Broad Institute to whole genome sequence 1000s of Mtb strains to create a catalog of mutations associated with drug resistance (DR) that will inform development of molecular diagnostics.

Central to the creation of the DR catalog are (1) whole genome assemblies that are as-complete-and-accurate-as-possible and (2) highly accurate polymorphism calls against a reference genome. For assembly, we use ALLPATHS-LG and Pilon to achieve nearly complete and highly accurate Mtb whole genome assemblies. First, using the reference assistance feature in the ALLPATHS-LG assembler and the complete H37Rv reference genome, we reduced the contig count of most Mtb assemblies by >50%. Greater improvement is observed when using more closely-related references for assisting. Second, using Pilon, Broad's assembly improvement tool, we can create more complete and accurate Mtb assemblies by correcting base errors and local mis-assemblies, extending contigs, and closing gaps. In one Mtb F11 assembly Pilon corrected 32 consensus errors, closed 7 gaps, and added 18kb of sequence.

For polymorphism detection, we rely heavily on Pilon to identify both small and large-scale events. While SNPs and small indels are generally straightforward to identify and multiple tools can reliably identify these small events, larger in-dels are more difficult to identify. Pilon, using only Illumina data, can capture insertion and deletion events of hundreds and thousands of bases in length. Further, Pilon can identify large-scale duplications and rearrangements, two event types found in previously sequenced Mtb strains.

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.:HHSN272200900018C

FF0152

**Genomics Capability Development and Collaborative Research with Global Engagement**

Helen Cui, Tracy Erkkila, Lance Green, Patrick Chain, Shannon Johnson, Cheryl Gleasner, Momchilo Vuyisich, Gary Resnick, Chris Detter
Bioscience Division, Los Alamos National Laboratory, NM 87544

**Contact:** Helen Cui, hhcui@lanl.gov, 505-665-1994

Genomics science and technologies are transforming research and development in many fields of life sciences globally. Los Alamos is assisting multiple countries and regions in advancing genomics capabilities, focusing on genomics scientific foundation, next generation sequencing technology, and analytics for pathogen detection and characterization and biosurveillance applications. Our current partner countries and regions include Jordan and Yemen in the Middle East and North Africa, Republic of Georgia, Kenya, Gabon, and South East Asia. We leverage our own long term research and development experience and expertise to assist the host nations to develop the capabilities that are urgently needed to address pressing challenges in infectious disease spreads, implementing safe laboratory practices, and developing infectious disease detection and characterization techniques that can be maintained and further developed by the host countries. Such collaboration efforts not only benefit the host countries and region in catching up with the state-of-the-art life science and technologies, but also build a trusted international community with shared passion in addressing global emerging infectious challenges and shared resources, essential for an effective global infectious disease surveillance approach and meeting International Health Regulation requirements.

FF0163

**Optimization of RNA-Seq to Determine Degradation Profiles and Identity of Forensically Relevant Samples**

Kate Weinbrecht (Oklahoma State University Center for Health Science, Tulsa, OK)

Research on the forensic applications of RNA analysis has increased greatly in the last decade. Many possible uses for RNA in forensic science have been indicated, including, use of RNA to identify specific tissues, monitoring degradation of RNA to determine time since deposition of a body fluid stain and post-mortem interval, and determining disease state, drug use, or mechanism of death through expression analysis of specific gene products. Ultimately, analyzing RNA from forensic samples can provide a wealth of information concerning when and how a crime occurred. Although recent research has indicated many possible forensic applications of RNA analysis, many questions remain concerning the behavior of RNA in degraded and limitedly available samples. In order for RNA to be used in regular forensic analysis, RNA degradation profiles and stable markers for tissue identity and degradation state must be established. Traditional molecular analysis techniques such as capillary electrophoresis and real-time PCR are capable of evaluating only a very minute portion of the transcriptome in a single assay. Next-generation sequencing of RNA allows for evaluation of the entire transcriptome of a sample in a single analysis run. With this research we plan to age forensically relevant samples, including blood, saliva, and semen, under controlled lab conditions for a known period of time. RNA obtained from minimally available and degraded biological stains is generally of low quality and low quantity. Optimization of sequencing library construction is being performed to obtain the most sequence information as possible from each degraded sample. RNA libraries are sequenced using the Ion Torrent PGM™ in order to establish transcriptome degradation profiles for each fluid type. Obtaining known transcriptome degradation profiles will allow for the selection of the most stable markers for tissue identity, sample degradation state, and approximate time since deposition of a biological stain.

FF0164

**Towards Improved NGS Workflows: Novel Method to Evaluate the Quality/Quantity of Genomic DNA**

Ken Taylor

Advanced Analytical Technologies, Inc.

Genomic DNA (gDNA) is a ubiquitous starting material for many molecular applications including Next-Generation Sequencing (NGS), which relies on the high-quality incoming gDNA of known concentration. Traditional approaches includes two separate instruments to fully characterize the gDNA, where the quality is assessed using agarose slab-gel electrophoresis and the quantity is measured using spectroscopic techniques such as UV-Vis or Fluorometry. In addition to being non-automated and labor intensive, agarose gel electrophoresis requires large amounts (nanograms) of samples due to low sensitivity, which may not be desirable when working with small amounts of precious samples (e.g. patient samples, FFPE tissues, single cell analysis, cancer heterogeneity etc). The traditional quantitation methods are non-automated and do not give information on the integrity of the gDNA. This necessitates a more sensitive, automated and increased throughput analysis technology to improve the NGS workflow. Reported herein are recent developments for qualifying and quantifying genomic DNA on a parallel capillary electrophoresis platform with LED-based fluorescence detection (*Fragment Analyzer*™). The system offers a more streamlined workflow, higher automation, lesser sample requirement and lower per run costs compared to traditional methods. This presentation will discuss the reproducibility and accuracy of the measurements of different forms of gDNA evaluated with the *Fragment Analyzer*™ and comparison of the results to traditional analysis methods. Advantages in terms of sensitivity, dynamic range, analysis time and RNA contamination will also be presented. Data analysis features that allow user-defined parameters to distinguish between different levels of degraded gDNA will be described. Results indicate good agreement with the expected concentration and sizing of genomic DNA. The *Fragment Analyzer*™ platform offers rapid and accurate quality/quantity check of genomic DNA on a completely automated, low-to-high throughput platform to augment NGS workflows for genomic research.

FF0167

**HGSC *i5K* Genomes Pilot Project**

Shannon Dugan, Jiaxin Qu, Yue Liu, Sandy Lee, Hsu Chao, Harsha Doddapaneni, Kim Worley, Donna Muzny, Richard Gibbs and Stephen Richards

*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030*

The goal of the i5k initiative is to sequence the genomes of 5,000 insect and related arthropod species over the next 5 years in order to provide a base reference for understanding arthropod evolution and phylogeny. Species to be included in this project were selected by scientists from around the world in order to address important medical, agricultural, ecological and scientific issues.

The BCM-HGSC is currently sequencing 31 arthropod genomes as a pilot project for this initiative. We have collected DNA and begun sequencing and assembly for insects with genomes ranging in size from 200Mb to 5Gb. In addition to genome size, differences in DNA homozygocity and other species specific features for these insects present difficult challenges.
Obtaining amounts of DNA needed to produce the Illumina fragment (180bp;500bp) and mate-pair (3Kb;8Kb) libraries as part of the project plan is further complicated by the size of the insect. Where inbreeding is not an option, we have tried to sequence and assemble genomes using a single individual in order to successfully drive the assembly process.

A project of this magnitude with such a large number of species to be completed requires not only low cost sequence generation, but also a method for high quality whole genome assembly.

For these insect species we are generating *de novo* assemblies using the ALLPATHS genome assembler followed by refinement using the genome assembly upgrade tools Atlas-link and Atlas-GapFill. We will perform automated Maker annotations, utilizing RNAseq from three different tissues to support that automated gene annotation. Initial sequence and assembly results from *Ceratitis capitata* (Mediterranean fruit fly) and *Athalia rosae* (turnip sawfly) are very promising with Contig N50s of approximately 50Kb. ScaffoldN50 values for these assemblies are 4.1Mb and 1.3Mb respectively. Additional strategies and assembly results will be presented.

FF0173a

**Strand-Specific High Throughput RNA Seq Library Preparation from Ovary Tissue RNA for Illumina Sequencing**

Monika Tomczyk, Jacob Berger, Paul Butler, Charles Troup, and Shanavaz Nasarabadi

IntegenX Inc., 5720 Stoneridge Drive, Suite 300, Pleasanton, CA 94588

Corresponding author: shanavazn@integenx.com

The ease of performing whole exome analysis with Next Generation Sequencing (NGS) technologies has become an invaluable tool in gene expression diagnosis of cancer cells. The cost benefits for determining regulatory changes incancer cells by whole transcriptome analysis has proven useful for tailoring treatment options for patients. An efficient, consistent, and reliable whole transcriptome library preparation method is necessary for successful transition of the NGS technologies to clinical applications. Due to the labile nature of mRNA and the tedium of processing large numbers of patient samples, automation of RNA-Seq libraries from total RNA would greatly improve the reliability and throughput of whole transcriptome library preparation.
We have explored the automation of whole transcriptome library preparation from total RNA of normal and cancerous ovary tissue. The poly A and ribosome depleted mRNA and whole exome libraries were prepared from total RNA on the Apollo 324™ system (IntegenX) followed by sequencing on the Genome Analyzer IIx (Illumina) While most conventional RNA-Seq library preparation methods convert mRNA to cDNA, our strand-specific library preparation method ligated the adapters directly to fragmented mRNA to preserve strand polarity. Preserving strand polarity of the transcript reduces the bioinformatics bottleneck. The Apollo 324System (IntegenX Inc.) was used to isolate poly A mRNA and ribosome-depleted mRNA from 6 to 48 samples of total RNA and prepare strand-specific mRNA libraries. The cDNA output was amplified in a bench thermocycler to yield whole transcriptome libraries in a single eight-hour work day. The DNA isolation from PCR product was performed on the Apollo 324. The libraries were normalized and sequenced on the GA IIx. A commercially available manual strand-specific library was used as the bench control for comparison and validation of our library preparations.

The gene expression profile of up-regulated and down-regulated genes was equivalent between the high-throughput libraries and the published data. The isolated poly A and ribosome depleted mRNA had an average of 0.3% rRNA contamination from 500 ng of total RNA. The automated RNA-Seq library preparation takes 8 hours for 48 samples, a 75% time saving compared to manual preparation methods. There was >90% correlation of gene expression between the bench RNA-Seq library preparation and the 48 RNA-Seq samples from the automated library preparation.

FF0173b

**Automated Directional Small RNA Library Preparation from Brain Tissue for Illumina GA Sequencing**

Jacob Berger, Paul Butler, and Shanavaz Nasarabadi

IntegenX Inc., 5720 Stoneridge Drive, Suite 300, Pleasanton, CA, 94568, USA

Corresponding author: shanavazn@integenx.com

Small RNA molecules play an important role in regulating gene expression, and identification of novel small RNA is being explored as a diagnostic and therapeutic tool for cancer. Small RNAs are uniquely challenging to sequence because they degrade faster than other RNA molecules, and represent a very small fraction (1% to 2%) of the RNA population depending on sample/tissue type. Although next generation sequencing allows researchers to rapidly sequence entire genomes and process many samples in parallel, the sample preparation can often be very tedious, time consuming, and prone to human error. We have explored novel methodologies for streamlining and automating directional library preparation processes while enhancing sequencing data quality.
In this study, we have demonstrated a process of enrichment for small RNA from raw tissue samples and automated library construction of the enriched small RNA to increase throughput and reduce human error. While conventional RNA-Seq methods does not permit directional sequencing of RNA, we developed a method of directional library preparation that has the advantage of preserving the strand polarity of the transcript to provide more valuable sequence data. Frozen rabbit brain tissue samples were pre-enriched manually using the miRVANA kit (Ambion) to purify and concentrate the miRNA. Using the PrepX™ RNA-Seq Library Kit for Illumina (IntegenX Inc.), eight small RNA libraries were prepared from as little as 10 ng of the enriched miRNA on the Apollo 324™ System in three hours.

Enrichment for small RNA increased the percentage of miRNA mapped reads from 54% to 74%, with over 65% of the reads mapping to known miRNAs. The percent mRNA mapped was reduced by 75% when compared to the non-enriched total RNA samples. The sequencing data for the enriched libraries gave twofold more miRNA reads than small RNA libraries prepared from non-enriched total RNA. However, the number of unique miRNAs was similar between the enriched and non-enriched samples.

FF0181

**Next Generation Sequencing, Tick Systematics and Functional Biology**

Mans BJ[1], de Klerk DG[1], De Castro MH[2], Pienaar R[1], Latif AA[1]

[1]Parasites, Vectors and Vector-Borne Diseases, Onderstepoort Veterinary Institute, Agricultural Research Council, South Africa

[2]Biotechnology Platform, Onderstepoort Veterinary Institute, Agricultural Research Council, South Africa

Ticks (Arachnida: Ixodida) are important vectors of a variety of pathogens. There are ~900 species that can be placed in three main families, hard ticks (Ixodidae), soft ticks (Argasidae) and the monotypic Nuttalliellidae. Current systematics uses mainly the nuclear 18S and 28S rRNA genes and mitochondrial genes. Next generation sequencing (Illumina) strategies were investigated to determine its usefulness for the assembly of these main markers and the discovery of new markers from fresh as well as museum specimens using genomic DNA sequencing. For this study, 25 different specimens (21 fresh, 4 museum) were investigated. Total genomic DNA yields varied from 2-450ng/ul. Samples were prepared using Nextera kits and sequenced using an Illumina HiScan. Data generated ranged from 0.5-4Gbp of ~100bp single read and paired end reads. Data were *de novo* assembled using CLC Genomics Workbench. The majority of assemblies yielded full-length 18S rRNA, 28S rRNA and mitochondrial genomes encoded in one contig. These generally corresponded with contigs that showed the highest coverage in the assemblies (37-2200 fold, depending on the sample). Assembly of fragmented *de novo* assembled contigs could be attained by mapping approaches. Coverage and assembly length for museum samples were generally better for mitochondrial genomes, while nuclear genes were more fragmented. Other highly abundant nuclear markers that could be identified in all samples included the core histone cassette (H4-H3-H2A-H2B). A next generation sequencing approach to identify markers for systematic analysis of single tick specimens is therefore a feasible alternative to labour some cloning and conventional sequencing approaches. Analysis of data derived from this project will also be analysed to identify potential candidate species for transcriptome and whole genome sequencing, with the aim of developing anti-tick vaccines and understand tick-host interactions better.

FF0201

**De novo and Hybrid Assembly for Challenging Bacterial Genomes**

Sagar M. Utturkar1, Dawn M. Klingeman2, Dale A. Pelletier1, 2, Christopher W. Schadt1, 2, Miriam L. Land2, Mitchel J. Doktycz1, 2 and Steven D. Brown1, 2

1) UT-ORNL Graduate School of Genome Science and Technology or Department of BCMB
2) Oak Ridge National Laboratory, Oak Ridge, TN, 37831

As part of the Plant-Microbe Interfaces (PMI) project 43 bacteria (33-69% GC content and 4-11 MB in genome size) that represent part of the Populus deltoides rhizosphere and root endosphere were isolated and their genomes were sequenced using an Illumina HiSeq2000 instrument. Quality based filtering and selection of appropriate Kmer values were found to be key steps for initial assembly. The best draft assemblies for 41 strains were recently published with an average contig number of 187. However, Rhizobium sp. strain CF080 and Burkholderia sp. strain BT03 were excluded from the publication due to high contig number (994 and 687 respectively). These strains were characterized by high percentage of transposable elements and repeat regions, which were not resolved by a single technology.

To improve the quality of assembly, additional sequencing was performed using the Roche 454, PacBio and Illumina mate-pair platforms for selected PMI isolates. Here, we describe hybrid assembly approaches that worked best for combinations of sequencing technologies. For a hybrid assembly that included 454 data, an initial Illumina assembly was shredded into 1.5 kb overlapping fragments and assembled with shotgun 454 data using Newbler version 2.6 generating 57 contigs for strain CF080. Addition of mate-pair data further reduced the number of contigs (36 for CF080). The ability of 3-5 kb long PacBio reads to generate closed genome structure is being evaluated. The PacBioToCA pipeline generated assemblies were comparable to 454 hybrids, if not better. The AHA method helped to merge the 454 hybrid assembly by aligning PacBio reads and generated scaffolds (9 for strain CF080). The PBJelly gap filling algorithm was employed to reduce the number of gaps within caffolds. This analysis confirms complimentary libraries and sequencing technology can be used to greatly improve genome assembly metric and provide genome sequences that are more amenable for manual finishing.

Acknowledgement:
This research was sponsored by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research, as part of the Plant Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov). Oak Ridge National Laboratory is managed by UT-Battelle LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. Travel support provided to Sagar Utturkar by the Graduate School of Genome Science and Technology, UTK.

FF0202

## *In silico* mRNA Transcriptional Profile Comparisons of *malus. x domestica* Fruit Peel and Pulp Tissues Using RNA-seq

Z. Chikwambi[*(1,2)], A. Christoffels[(2)], D.J.G Rees[(1)]
[(1)] Biotechnology Platform, Agricultural Research Council, Private Bag X5, Onderstepoort, 0110, South Africa
[(2)] University of the Western Cape, Private Bag X17, Bellville, Cape Town, South Africa

The growth and development of 'Golden delicious' fruit occurs over a period of about 135 to 150 days after anthesis (daa) to fully ripe. During this period morphological and physiological changes occur resulting in defined fruit size and quality. These changes are a result of spatial and temporal differential gene expressions in 'Golden Delicious' fruit defining the properties of the fruit peel and pulp tissues as well as the acceptance of the fruit on the market. Peel and pulp tissue-up-regulated transcripts were spatially and temporally enriched for pathways defining tissue specific properties. Photosynthesis, carbon fixation, fatty acid metabolism, terpenoid, carotenoid and flavonoid pathways encoding transcripts were mostly enriched in the peel tissue. On the other hand oxidative phosphorylation, plant hormone signal transduction and $C_4$-carbon fixation pathways were enriched in the pulp tissue. This study has thus provided an in depth catalogue and elucidation of the pathways enriched in peel and pulp 'Golden Delicious' tissues, which might function in defining the properties of the tissues, and their contribution thereof to total fruit development. An in depth understanding of the molecular mechanisms of fruit development and fruit tissue contribution to development allows for refined and focused cultivar breeding and manipulation approaches.

**Key words:**
KEGG pathway enrichment, apple fruit peel, apple fruit pulp, 'Golden Delicious' fruit development
Talk

- Corresponding author: Z.Chikwambi,
- Email: zchikwambi@gmail.com, chikwambiz@arc.agric.za
- Tel: +27 21 529 9121

FF0234

**Sequencing Biases of Next-Generation Sequencing Platforms**

K. Davenport, M. Scholz, H. Daligault, O. Chertkov, H. Teshima, C. Munk, S. Feng, T. Freitas, T. Erkkila and P. Chain

Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory, Los Alamos, NM.

One of the greatest barriers to assembling a genome is the result of variable coverage-specifically, dramatically low coverage of regions of the genome resulting from random shotgun sequencing. To improve coverage, close gaps and correct misassemblies resulting from low coverage, we utilize two sequencing platforms with different strengths and weaknesses: Illumina (short reads with a very low sequencing error rate) and Pacific Biosciences (long reads with a significantly higher error rate). In an effort to determine the role of GC content or sequence motifs in coverage variation, we have analyzed both platforms' coverage of completed bacterial genomes in relation to varying GC content. Variability of coverage was also compared between members of the same species as well as within the same genus.

LA-UR-13-23468

FF0235

## Gel Microdroplets with Single Cell Genomics Obtains Complete Genomes from the Human Microbiome: Potential for Broad Applications

Armand E Dichosa[1], Michael S Fitzsimons[3], Patrick S Chain[1], Shawn S Starkenburg[1], J Chris Detter[2], and Cliff S Han[1]
[1] Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory, Los Alamos, NM 87545. [2] B-DO: Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545. [3] NuGen Technologies, San Carlos, CA 94070

The collective bacteria, archaea, and viruses that comprise the many facets of the human microbiome significantly impact the overall health of the host. Equally important to understanding the host-microbe interactions is the microbe-microbe dynamics. Specifically, genetic exchange via natural transformation, transduction, and conjugation allows for inter/intra-species transfer and integration of genetic material, thereby altering the microbe's overall functional activities and fitness through variations of individual genomes. The challenge, therefore, is to obtain multiple, [near] complete genomes from representative cells for in-depth comparative genome analyses. As single cell genomics (SCG) offers rapid, culture-independent means to determine both the taxonomic identity and potential physiology from the genomic perspective, obtaining the complete genome from an isolated single cell remains elusive. Our prior work has demonstrated that multiple, clonal genomic templates (via whole cells) greatly improve genomic assemblies. However, as a vast majority of cells cannot grow in culture, possibly due to necessary growth signals and/or specific co-microbial influences, obtaining substantial genomic template is impossible.

To simultaneously overcome the amplification bias inherent in multiple displacement amplification (MDA) and to broadly enrich for cells under more favorable growth conditions, we hypothesized that capturing single cells in gel microdroplets (GMD; one cell in each GMD), for growth among its native community will yield clonal microcolonies sufficient to achieve complete genomes.

We tested our hypothesis on human oral and gut microbiomes by singly capturing cells in agarose-based GMD spheres (~40 μm) for *in vitro* co-cultivation, respectively. Cell-to-cell communication and nutrient exchange occurs during co-cultivation due to the porous properties of each GMD. Thus, single cells form clonal microcolonies conveniently packaged in each GMD, which was amenable for subsequent MDA. Our results yielded near complete genomes from several oral and gut microbes, which provided remarkable insight into intragenomic species variation from the *Streptococcus* oral inhabitants, which was likely due to homologous recombination, and little genomic variation from the *Enterococcus* gut inhabitants. Significant regions in the oral *Streptococcus* genomes were highly conserved and are likely essential to growth under our experimental conditions, while the regions containing more differences were involved with pathogenicity and energy metabolism.

Our findings show how significantly active (or inactive) the members of microbiomes are in recombining specific segments of DNA with each other. Our work also raises questions as to how we identify a "species", since completed genomes provides more taxonomic resolution than the standard 16S rRNA phylotyping. These and other insights about the biology and functional roles bacteria play in their respective microbiomes are revealed from the perspective of complete genomes, which could not have been achieved without the use of GMD.
As this study demonstrates the benefits of GMD with SCG for human microbiome studies, our team envisions its potential for broad applications involving bioenergy research, host-pathogen / cell-cell interactions, and environmental community studies.

FF0236

**Gene Expression Differences between Intracellular and Extracellular Yersinia Pestis using Rna-seq**

Bin Hu1, Sofiya N. Micheva-Viteva2, Momchilo Vuyisich1, Vladimir L. Motin3, Elizabeth Hong-Geller2, and Patrick S. Chain1

1. Bioenergy and Biome Sciences, Bioscience Division, M888, Los Alamos National Laboratory, Los Alamos, NM 87545 2. Biosecurity and public health, Bioscience Division, M888, Los Alamos National Laboratory, Los Alamos, NM 87545 3. Department of Pathology, University of Texas Medical Branch, Galveston, TX 77555

Yersinia pestis, the causative agent of plague, is known to dramatically alter gene expression during its transition from the flea to mammalian host. Given recent evidence that Y. pestis can replicate within host cells, we investigated gene expression differences between intracellular Y. pestis cells isolated from infected host cells compared to pathogen in the extracellular milieu. In this study, we infected the human macrophage cell line THP-1 with virulent Y. pestis CO92 and extracted RNA from the intracellular and extracellular pathogen fractions for rRNA depletion and RNA-seq transcriptome profiling with RNA-seq. Results were compared to control RNA from Y. pestis grown in media alone. Our studies revealed approximately 2900 Y. pestis genes with significant differential expression (padj < 0.05) upon infection of THP-1 cells. We found that 1760 genes exhibited significant expression differences (padj < 0.05) between intracellular and extracellular Y. pestis. Of particular interest, 186 out of the 1760 genes were only differentially regulated within host cells and were not found to be regulated between extracellular and the control conditions. Several known virulence genes were found among these 186, including components of the Type III secretion system, YscS, YscT, and YscF, and genes involved in iron transport. Other genes uniquely induced within the intracellular environment encode several metabolic enzymes, ABC-transporters, and proteins of unknown functions. The unique gene expression profile of intracellular Y. pestis during infection suggest that different virulence mechanisms function during intracellular invasion of the host.

FF0239

**Whole-genome Based SNP Phylogeny**

Sanaa Ahmed

Los Alamos National Lab, Genome Science Group

Single nucleotide polymorphisms (SNPs) are the most abundant phylogenetically informative form of genetic variation among closely related microbial species, strains and isolates. SNPs can confer selective advantage for microbial organisms competing for finite resources and can thus serve as powerful genetic markers for distinguishing phylogenetically closely related strains. To facilitate rapid SNP discovery in microbial genomes, LANL has developed an application, for genome-wide identification and characterization of SNPs, be they from raw data (sequencing reads) or assemblies (draft or completed genomes). Current methods only allow partial genome-wide SNP identification, and only focus on draft, or complete genomes. Our software is unique as it can identify SNPs from reads and combine these data with SNPs from any set of genomes or contigs. This method is rapid and extensible, building upon the multiple sequence alignment for any set of organisms already studied. Two empirical examples using >200 genomes available that are within either the *Burkholderia* or *Escherichia* lineage, show rapid and robust clustering even when using more distant, outgroup clades.

FF0240

**LANL Genome Science Program Sequencing Capabilities and Projects**

Hong Shen

Los Alamos National Laboratory, Genome Science Group

The Genome Science Program of the Bioenergy and Biome Sciences group (B-11) in Bioscience Division at Los Alamos National Laboratory (LANL) specializes in high-throughput genomics and genome analysis for a variety of internal and collaborative projects. This includes draft sequencing, genome improvement, transcriptome and metagenome sequencing. Sponsored projects are in support of DOD, DHS, DOE and national security missions. We have active projects in many areas, including pathogen biology, biosurveillance, energy, and bioremediation. The LANL Genome Science Program currently utilizes a number of platforms that include traditional capillary Sanger Sequencing, 454 pyrosequencing, Illumina sequencing (Genome Analyzer, Hiseq 2000, Miseq), PacBio single molecule real-time sequencing (PacBio RS), and ion semiconductor sequencing (Ion Torrent PGM and Ion Proton). Depending on the project, data is analyzed in a highly automated fashion with dedicated analysis pipelines, or data are custom analyzed by bioinformaticists.

FF0241

Tracey Freitas

Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory

Characterizing the complexity and relative abundance of microbes in metagenomic samples relies on the assignment of sequencing reads to a reference set of sequences in some form or another. Attempts have been made to unambiguously assign reads to a reduced reference set to reduce or eliminate false positives, however the most successful approaches currently incorporate only those regions originating from gene sequences. We have created a set of signature sequences based off the entire length of the bacterial reference genomes, reducing the dataset by 35%, and profiled the Human Microbiome Project's Mock samples with the open-source aligner BWA against our database. Using this reduced reference set, we were able to obtain taxonomic-specific (Strain, Species, and Genus) profiles in under 10 minutes each, identifying the 20 completed bacteria genomes present within the sample. We show the effect that the trimmed read length has on false positives and that allowing errors of up to 2 bases and 1 gap do not significantly affect the results. In addition, we explore the more difficult scenarios of assigning taxonomies to (i) synthetic genomes, (ii) novel genomes and (iii) an air filter metagenome spiked with a known pathogen

FF0242

**High quality sequencing and assembly of bacterial genomes using NEBNext reagents**

Momchilo Vuyisich*, Ayesha Arefin, Karen Davenport, Shihai Feng, Cheryl Gleasner, Kim McMurry, Jennifer Price, Matthew Scholz, and Patrick Chain.

Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory
*Corresponding author: vuyisich@lanl.gov

Shotgun sequencing of bacterial genomes has traditionally required large amounts of genomic DNA (~1 □g). The ne much smaller amounts of input genomic DNA. We have evaluated the utility of NEBNext Ultra for resequencing and *de novo* assembly of four bacterial genomes, and compared its performance with the TruSeq library preparation kit. The NEBNext Ultra reagents enable high quality resequencing and *de novo* assembly of a variety of bacterial genomes when using 100 ng of input genomic DNA. For two bacterial genomes with the highest GC content (*Burkholderia* spp.), we also show that the quality of both resequencing and *de novo* assembly is not decreased when only 10 ng of input genomic DNA is used.

FF0243

**DynaTrim, a dynamic trimming method for next generation sequencing Data**

Shihai Feng, Chien-Chi Lo, Matthew B Scholz, Paul Li and Patrick SG Chain*

Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory

We developed a dynamic trimming method, DynaTrim, that utilizes both window-based trim and single-based methods. Moreover, our method does not require a single user defined quality threshold, which could be not appropriate over the entire read length and among different sets of sequencing reads. Our method utilizes the quality distribution of entire sequence samples at given position and its adjacent reads qualities. It greatly improves the true positive rate similar false positive rate of finding errors. It is independent of computational platforms and sequencing technologies.

FF0244

**Detection Of A Low Abundance Pathogen In A Mixed Community Sample By Use Of Multiple Sequencing Platforms**

Matthew Scholz

Genome Science Programs, Biosciences Group B-11, Los Alamos National Laboratory

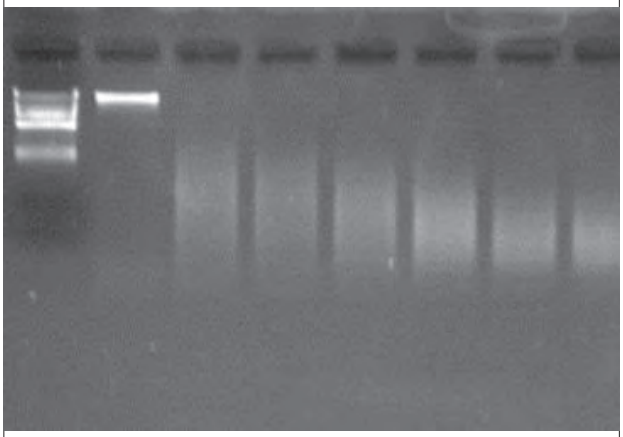
Identification of a single organism from a community sample by next generation sequencing (NGS) is a difficult process, complicated by the lack of genome references for many environmental organisms. A series of sequencing and analysis tasks were performed to determine the ability of NGS technologies to detect a pathogen from a mixed microbial community. To this end, an air filter sample, was spiked with a low copy number of a foreign pathogen (Francisella tularensis), DNA was extracted, and sequenced using multiple NGS platforms (HiSeq, PacBio, 454 GLS). . DNA extracts from this sample were used to prepare sequencing libraries for 454, Illumina, and PacBio sequencing (with or without prior amplification), and reads from each platform were analyzed both separately and in a combined fashion to evaluate the community profiles derived. . Additional effort produced a community profile from the remaining reads to classify the other organisms present in the sample The relative abundance of F.tularensis in the sample was very low, and determined to make up between 0.005-3.0% of the total genetic material sequenced. Due to the low abundance of Francisella, only Illumina HiSeq generated sufficient depth of coverage to allow definitive identification and characterization of the target genome to the nearest neighbor species, however sufficient read depth was achieved for all technologies surveyed to allow confirmation of the identity of the pathogen. Reads were assembled, and both reads and assembled contigs were utilized to analyze genetic differences compared with the nearest reference strain. A total of 43 SNPs were identified to the reference genome, of which 11 were annotated as non-synonymous SNPs. To characterize the remaining reads, 454 and a random selection of illumina libraries were used. MEGAN analysis indicates that the most abundant families present on the filter were Psudomonadales, Xanthomonadales, and Rhizobiales.

# *Poster Session Notes*

# Check quantity AND quality of gDNA with one instrument.

## The Fragment Analyzer™ Automated CE System

### The Past

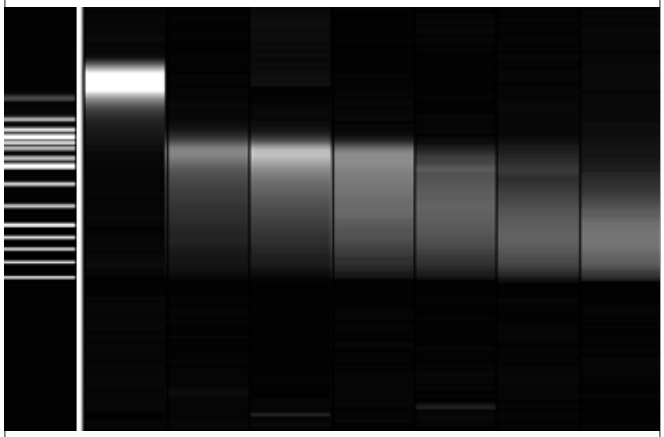


*Human genomic DNA. Traditional manual agarose slab gel shows intact gDNA in the second lane. Gel images in remaining lanes show varying levels of gDNA degradation.*

### Fragment Analyzer™ Benefits

- **No more pouring gels.** Automated simultaneous analysis of 12 or 96 samples.
- **Higher sensitivity than agarose gels.** Use small amounts of gDNA samples. (0.1 ng)
- **Ultra fast lower marker** (set to 1bp) migrates faster than degraded gDNA for superior quality and quantity assessment.
- **Good sizing capability** to differentiate degraded, partially degraded or intact gDNA.
- **See RNA contamination** in gDNA extractions.

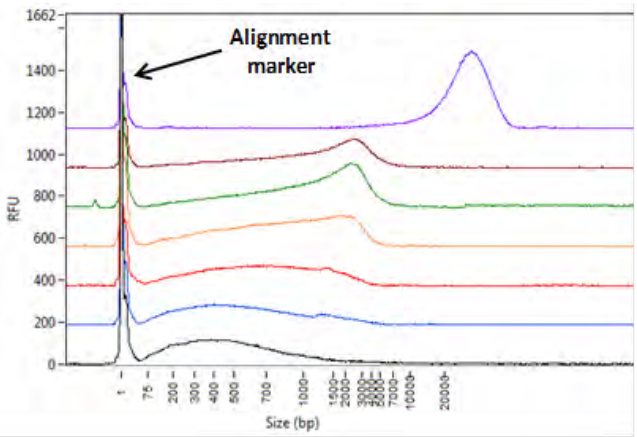
### The Future



*Same sample of human gDNA, identical results.*

*BELOW: Raw data is captured by automated capillary electrophoresis system, as seen in electropherogram overlay. >20,000 bp peak indicates intact gDNA on the upper-most trace.*

*ABOVE: Data can then be processed and presented in a variety of ways, such as this digital gel image.*




ADVANCED ANALYTICAL

Phone: 515-296-6600 | www.aati-us.com

# *Poster Session Notes*

| *05/30/2013 - Thursday* | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| **7:30 - 8:30am** | **Breakfast** | **x** | **Santa Fe Breakfast Buffet** | **Sponsored by NEB** |
| **8:30 - 8:45** | Intro | **x** | Welcome Back | **TBD** |
| **x** | Session Chair | **x** | Session Chairs | Chair - Mike Fitzgerald<br>Chair - Nadia Fedorova |
| **8:45 - 9:30** | **Keynote** | **FF0121** | **Challenges and Opportunities in Strain-level Comparative Genomics** | **Mark Adams** |
| **9:30 - 9:50** | Speaker 1 | **FF0210** | Genome of the Pathogen Porphyromonas gingivalis Recovered from a Biofilm in a Hospital Sink using a New Platform and the Single-cell Assembler SPAdes | **Glenn Tesler** |
| **9:50 - 10:10** | Speaker 2 | **FF0139** | Characterization of the Culex Mosquito Virome in California by Metagenomic Sequencing | **Stanley Langevin** |
| **10:10 - 10:30** | Speaker 3 | **FF0205** | Food Security: the 100K Pathogen Genome Project | **Bart Weimer** |
| **10:30 - 10:50** | **Break** | **x** | **Beverages and Snacks Provided** | **Sponsored by OpGen** |
| **10:50- 11:10** | Speaker 4 | **FF0152** | Genomics Capability Development and Collaborative Research with Global Engagement | **Helen Cui / Tracy Erkkila** |
| **11:10 - 11:30** | Speaker 5 | **FF0010** | Establishing Regional NextGen Whole Genome Sequencing | **Adam Kotorashvili** |
| **11:30 - 11:50** | Speaker 6 | **FF0194** | Biobanking and Metagenomics Platforms for Pathogen Discovery | **George Michuki** |
| **11:50 - 1:10pm** | **Lunch** | **x** | **New Mexican Lunch Buffet** | **Sponsored by Perkin Elmer** |
| **x** | Session Chair | **x** | Session Chairs | Chair - Donna Muzny<br>Chair - Johar Ali |
| **1:10 - 1:30** | Speaker 7 | **FF0071** | The $1M Metagenomics Algorithm Challenge | **Edward Wack** |
| **1:30 - 1:50** | Speaker 8 | **FF0237** | Analytical Process for Interactive Analysis of Deep Sequencing Data on a Laptop | **Ben McMahon** |
| **1:50 - 2:10** | Speaker 9 | **FF0160** | Metagenomic Applications for Diagnostics and Etiologic Agent Discovery | **Joe Petrosino** |
| **2:10 - 2:30** | Speaker 10 | **FF0238** | Clade-specific Genomic Signatures as a Method for Accurate Profiling of Metagenomic Datasets | **Patrick Chain** |
| **2:30 - 2:50** | Speaker 11 | **FF0117** | Metagenomic and Metatranscriptomic Analyses of the Complex Community of a Tropical Wastewater Treatment System | **Stephan Schuster** |
| **2:50 - 3:10** | **Break** | **x** | **Beverages and Snacks Provided** | **Sponsored by CLCbio** |
| **3:10 - 5:00pm** | **Tech Time Talks (15 min each)** | **FF0051b** | CLC bio Products and Supported Applications | **Marta Matvienko** |
| | | **FF0075** | The CLC Microbial Genome Finishing Module | **Martin Simonsen** |
| | | **FF0030** | Latest Advances in Bioinformatics Computing | **George Vacek** |
| | | **FF0149** | Implementing Fast Sequence Analysis Tools Using a Cray XMT2 | **Sterling Thomas** |
| | | **FF0221** | Genomics Applications in the Cloud with the DNAnexus Platform | **Andrey Kislyuk** |
| | | **FF0078** | Sequence Consensus Algorithms & Hierarchical Genome Assembly Process for Effective De Novo Assembly with SMRT® Sequencing | **Aaron Klammer** |
| | | **FF0126** | Haplotype Assembly Refinement & Improvement | **Christine Olsen** |
| **5:30 - 8:00pm** | **Happy Hour** | **x** | **Happy Hour at Cowgirl Cafe - Sponsored by LifeTech<br>Map Will be Provided** | **Sponsored by LifeTech** |
| **8:00 - bedtime** | **on your own** | **x** | **Dinner and Night on Your Own - Enjoy!!!** | **x** |

# *NOTES*

# *Speaker Presentations (May 30th)*

Abstracts are in order of presentation according to Agenda

Keynote

FF0121

**Challenges and Opportunities in Strain-level Comparative Genomics**

Mark D. Adams, J. Craig Venter Institute, San Diego, CA

Continued declines in the cost of draft genome sequences has meant that it is feasible to sequence many closely related strains of a given organism – from microbes to humans. By linking genotypes to phenotypes, it is possible to conduct a true "genome-wide" association study to define the genetic basis of key traits including antibiotic resistance in bacteria. For many genomes, however, interesting genetic features are disproportionately located in difficult to assembly genome regions such as duplications, repetitive sequence, plasmids, etc.

We have explored the diversity of closely related strains of *Acinetobacter baumannii* and *Klebsiella pneumoniae* to characterize the genetic context of laterally transferred antibiotic resistance genes and the extent of lateral transfer. Inference of the core phylogeny was based on high quality SNVs across all strains. A pangenome analysis pipeline was developed to identify core and accessory genes and link gene content with phenotypic information.

Within a single MLST group, there is considerable diversity, particularly of plasmid sequences and the location of insertion sequences. Single molecule sequencing using Pacific Biosciences SMRT system enabled up to complete chromosome length contigs to be assembled and clarified the structure of plasmids.

# *NOTES*

FF0210

**Genome of the Pathogen Porphyromonas gingivalis Recovered from a Biofilm in a Hospital Sink using a New Platform and the Single-cell Assembler SPAdes**

Glenn Tesler

University of California, San Diego Department of Mathematics

The lion's share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using conventional technologies. Shotgun metagenomics is an alternative, but is limited in its ability to detect strain variations. In this study, we present a single-cell genome sequencing approach to address these limitations and apply it to bacterial cells within a complex biofilm from a hospital bathroom sink drain. Single-cell genomics has been applied to capture novel genomes in marine and soil environments, and now, we show its application in a healthcare facility. A newly developed, automated platform was used to generate genomic DNA by the multiple displacement amplification (MDA) technique, revealing a broad range of bacteria covering 25 different genera. Here we focus on recovery of a novel strain of Porphyromonas gingivalis (P. gingivalis JCVI SC001) using the single-cell assembly tool SPAdes. Assembly of single-cell data is challenging because of highly non-uniform read coverage as well as elevated levels of sequencing errors and chimeric reads. SPAdes is a new assembler for both single-cell and standard (multicell) assembly. SPAdes provides information about genomes of uncultivatable bacteria that vastly exceeds what may be obtained via traditional metagenomics studies.

Software: http://bioinf.spbau.ru/spades

Articles:
McLean et al. (2013).
Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform.
Genome Research, published online ahead of print on April 5, 2013.
doi:10.1101/gr.150433.112

Bankevich and Nurk et al. (2012).
SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5): 455-477.
doi:10.1089/cmb.2012.0021

FF0139

**Characterization of the *Culex* Mosquito Virome in California by Metagenomic Sequencing**

S.A. Langevin[1], O.D. Solberg, D.J. Curtis[1], S.S. Wheeler[2], Z.W. Bent[1], V.A. Vandernoot[3], W.K. Reisen[2], and T.W. Lane[1].

[1]Systems Biology Department, Sandia National Laboratories, Livermore CA, USA. [2]Center for VectorBorne Diseases, Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis CA, USA. [3]Biotechnology and Bioengineering Department, Sandia National Laboratories, Livermore CA, USA.

Monitoring infections in vectors such as mosquitoes, sand flies, and ticks to identify human pathogens may serve as an early warning detection system to direct local government preventive measures (vector-control, public outreach). One major hurdle in mosquito-borne virus detection assays is the ability to screen large numbers of vectors for human pathogens without the use of genotype-specific molecular techniques. Current vector surveillance programs test for only a subset of pathogens known to circulate in a given geographic area and do not have the diagnostic capacity to screen for novel/emerging public health threats. Metagenomic sequencing provides an unbiased platform capable of identifying known and unknown pathogens circulating within a vector population, but utilizing this technology is time-consuming and costly for vector-borne disease surveillance programs. We developed novel metagenomic sequencing protocols to generate Illumina® compatible libraries and remove abundant host genetic background to enrich for microbial sequences circulating in vectors at the population level. This cost-effective biosurveillance approach was implemented to characterize microbial populations circulating in various *Culex species* mosquitoes throughout California during the 2012 West Nile virus outbreak. We identified novel and well-documented mosquito-borne viruses (Bunyaviridae, Flaviviridae, Mesoniviridae, Parvoviridae, Picornaviridae, Rhabdoviridae, and Reoviridae) present in spatially/temporally defined *Culex* populations and determined their phylogeny. These findings demonstrate metagenomic sequencing as an effective surveillance tool to screen vector populations for human pathogens.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

FF0205

**Food Security: the 100K Pathogen Genome Project**

Bart Weimer

UC Davis

Food safety and security is an increasingly important public health concern. As relatively few organisms continue to cause most of the associated disease the ability to more rapidly detect and curtail outbreaks is critical to improve public health. While many routine analytical methods are legislated for use with food adoption of NGS tools is providing new and faster methods to find outbreaks. Lack of a reference database of organisms is hindering the progress to make analytical decisions with finely resolved information. Use of SNPs is providing a method to determine the occurrence of strains that are associated with disease. The 100K pathogen genome project is sequencing 100,000 bacterial pathogens associated with food and zoonotic disease. This will provide a reference database to compare for public health, but as important it will also define the pan-genomic space for many of the leading pathogens. The public 100K bioproject database at NCBI will also provide enough genome sequence to enable the genomic community to determine fundamental questions associated with bacterial evolution, gene exchange, and data to provide accurate bioclock measurements for robust biomarker discovery. The project will also produce 1000 closed genomes for use in metabolic and epigenetic questions.

FF0152

**Genomics Capability Development and Collaborative Research with Global Engagement**

Helen Cui, Tracy Erkkila, Lance Green, Patrick Chain, Shannon Johnson, Cheryl Gleasner, Momchilo Vuyisich, Gary Resnick, Chris Detter
Bioscience Division, Los Alamos National Laboratory, NM 87544

**Contact:** Helen Cui, hhcui@lanl.gov, 505-665-1994

Genomics science and technologies are transforming research and development in many fields of life sciences globally. Los Alamos is assisting multiple countries and regions in advancing genomics capabilities, focusing on genomics scientific foundation, next generation sequencing technology, and analytics for pathogen detection and characterization and biosurveillance applications. Our current partner countries and regions include Jordan and Yemen in the Middle East and North Africa, Republic of Georgia, Kenya, Gabon, and South East Asia. We leverage our own long term research and development experience and expertise to assist the host nations to develop the capabilities that are urgently needed to address pressing challenges in infectious disease spreads, implementing safe laboratory practices, and developing infectious disease detection and characterization techniques that can be maintained and further developed by the host countries. Such collaboration efforts not only benefit the host countries and region in catching up with the state-of-the-art life science and technologies, but also build a trusted international community with shared passion in addressing global emerging infectious challenges and shared resources, essential for an effective global infectious disease surveillance approach and meeting International Health Regulation requirements.

FF0010

**Establishing Regional NextGen Whole Genome Sequencing**

Adam Kotorashvili[1], Jason Farlow[1], Cheryl Gleasner[2], Lance Green[2], Tracy Erkkila[2], Ben Allen[2], Chris Detter[2]

Richard G Lugar center for Public Health Research[1]
Los Alamos National Laboratory[2]

Georgia is a small former Soviet Union country located in eastern part of Europe with strategic location between black and Caspian seas in Caucasus region. There are two more other former Soviet countries in this region: Azerbaijan and Armenia. In the region there are many places for endemic disease such as Anthrax, Tularemia, plague and so on, which means that region is very interesting science wise.

The Richard G Lugar Center for Public Health Research (CPHR) is a joint project of the U.S and Georgian governments and was constructed with funds allocated by the U.S. Department of Defense to provide an early warning system for occurrence and potential spread of infectious diseases affecting Georgia and the rest of the world. Establishment of the new central laboratory in Tbilisi and its associated satellite laboratories in the region represents a unique opportunity to provide state of the art advances in life sciences that will assist health professionals of the region in combating infectious disease pathogens of concern. With this system in place and operational, Georgia is now able to meet reporting requirements under the WHO International Health Regulations (IHR) and the World Organization for Animal Health (OIE), as well as support the One Health Initiative. CPHR is an important resource for meeting the challenge of public and animal health improvement and can serve as an economic and intellectual driver for science in Georgia. The Genome Center at CPHR provides a unique opportunity not just for Georgia but for whole region in applied genomics. The facility houses a 3130xl genetic analyzer and MiSeq Illumina platform. We have got MiSeq about two month ago and we have already sequenced several viral Tb phage genomes. There is huge interest of Genome Center not just from Georgian universities, hospitals or private labs but also from other the countries in the Region: including Azerbaijan and Armenia, as well as Turkey and Ukraine.

FF0194

**Biobanking and Metagenomics Platforms for Pathogen Discovery**

George Michuki[*], Absolomon Kihara[*], Alan Orth[*], Cecilia Rumberia[*] and Steve Kemp[*]
[*] International Livestock Research Institute

There has been documented increase of emergence and re-emergence of new zoonotic infectious diseases in Africa in the past years. The pathogen discovery team at the International Livestock Research Institute (ILRI) addresses the problem in Kenya and beyond. The genomics platform at ILRI which is an integration of the biorepository (Biobank), first and second generation sequencing platforms and high performance computing systems facilitates the pathogen discovery work.

The biorepository contains samples that are safely stored in ultra low temperature freezers (LN2 Tanks), are uniquely barcoded, are easily traceable using a LIMS system and have rich metadata. The material metadata is constantly updated with results from the analysis that is conducted on them. The biobank is under real time security surveillance and temperature monitoring systems. The biobank provides a rich source of samples and materials to support the work that is conducted by the sequencing platform. In addition to pathogen discovery, the biorepository provides a platform for genetic conservation by providing a large pool of materials from different diversities that can be used to recreate wiped out populations.

The sequencing platform is equipped with capillary sequencers (ABI 3130xl, ABI 3730xl and ABI 3500xl), the Roche 454 GSFLX genome sequencer and Illumina MiSeq and backed by level II and III laboratory facilities. The high-performance computing environment supports the bioinformatics analysis and storage of generated data using 88 compute cores and 31TB of network-attached GlusterFS storage. Scheduling of jobs to batch and high-memory compute nodes is managed by the SLURM resource manager.

The genomics platform has supported both local and international institutions generating outputs ranging from whole/partial genome sequences of Rift Valley Fever, Equine Encephalosis, Blue Tongue, African Swine Fever, Ndumu, Semliki forest, Dugbe, Bunyamwera, Newcastle diseas and Pigeon Paramyxo - viruses among other organisms including, bacteria, plants and animals. A number of the outputs are also available to the public on NCBI database.

# *NOTES*

# *NOTES*

# Lunch

**11:50 – 1:10pm**

## Sponsored by

# *Notes*

FF0071

## The $1M Metagenomics Algorithm Challenge[1]

Edward C. Wack[a], Darrell O. Ricke[a], Anna Shcherbina[a], C. Nicole Rosenzweig[b], Jessica Hill[b], Lee Ann McCue[c], Rachel Bartholomew[c], Darren S. Curtis[c], Aaron R. Phillips[c], Christian Whitchurch[d]

The Defense Threat Reduction Agency (DTRA) is sponsoring a public competition to acquire a revolutionary new software capability to analyze genomic data derived from complex biological samples. With the advent and proliferation of high throughput, same-day DNA sequencing capability, downstream analysis and interpretation is the rate-limiting step to action. Annotation and interpretation of data from complex clinical and environmental samples, with dozens to hundreds of organisms, are particularly challenging. A team from MIT Lincoln Laboratory, Edgewood Chemical and Biological Center, and Pacific Northwest National Laboratory has developed and is executing the competition on behalf of DTRA. The competition will result in an algorithm that rapidly and accurately characterizes complex DNA samples for constituent organism strain-level identification, genes and variants. This presentation will discuss the structure of the competition, goals and criteria, current status, evaluation plans and lessons learned.

[a] MIT Lincoln Laboratory, 244 Wood St., Lexington MA 02420, [b] Edgewood Chemical Biological Center, 5183 Blackhawk Road, Bldg 3150, Aberdeen Proving Ground, MD 21010, [c] Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA 99352, [d] Defense Threat Reduction Agency, 8725 John J Kingman Rd #6201 Fort Belvoir, VA 22060

[1] This work is sponsored by the Defense Threat Reduction Agency under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and not necessarily endorsed by the United States Government.

FF0237

**Analytical Process for Interactive Analysis of Deep Sequencing Data on a Laptop**

Ben McMahon, Judith Cohn, Joel Berendzen Cathy Cleland, Judith Cohn, Mira Dimitrijevic, Nick Hengartner

Theooretical, Bioscience, Physics, and Computer Science Divisions, LANL.

The combination of modern information science tools, evolutionary theory, and the extensive database of reference genomes has enabled a new class of algorithms to rapidly organize, annotate, and understand deep sequencing data, with computational power readily available in laptop or desktop computers. We present an analytical protocol which shows how signature-based analysis with Sequedex (sequedex.lanl.gov, sequedex.readthedocs.org/en/latest/), which was developed to understand soil metagenomes (Berendzen, et al. BMC Res. Notes, 5:460, 2012) can be combined with read-mappers, de-novo assemblers, sequence-aligners, phylogenetic tree builders, and statistical algorithms to analyze phylogenetic and functional profiles. Together, this analytical process enables interactive interrogation of 100 million DNA or transcriptomic reads to answer a wide variety of biological questions of a sample. Iterative improvement of peptide signature annotation promises even greater value in the future, as information from a variety of 'omics based capabilities becomes available.

FF0160

**Metagenomic Applications for Diagnostics and Etiologic Agent Discovery**

Matthew C. Ross[1], Ginger A. Metcalf[2], Embriette Hyde[1], Khanh Thi-Thuy Nguyen[3], David Wheeler[2], Rodrigo Hasbun[4], Joseph F. Petrosino[1,2]

[1]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX; [2]The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; [3] The University of Texas MD Anderson Cancer Center, Houston, TX; [4]University of Texas Health Science Center-Houston, Department of Infectious Disease

Historically, identification of viral and bacterial etiologic agents has relied on propagation of the suspected agent in the laboratory. More recently, advances in molecular biology have enabled researchers to characterize a subset of newly discovered viruses and bacteria, but these methods require some knowledge of the agent being sought. There are numerous diseases whose epidemiology suggests an infectious cause, but no etiological agent can be identified. We are utilizing both bacterial and viral metagenomics in an attempt to identify agents associated with diseases where an infectious agent is suspected. These include diseases such as Type 1 Diabetes (examining >18,500 samples with the NIDDK TEDDY and JDRF nPOD cohorts), aseptic meningitis (where typical etiologic agents are not able to be detected using traditional diagnostics), and unknown agents of sepsis in immunocompromised subjects (e.g. those with HIV or those undergoing chemotherapy).

Metagenomics strategies for etiologic agent discovery and diagnostic detection has multiple theoretical advantages over past methods: it is highly-sensitive; able to detect sub-PFU levels of viruses in spiked clinical samples, there is no requirement for propagation of the suspected agent in the laboratory, and there is no need for prior knowledge of the agent to be detected. For suspected bacterial agents we are using 16S rRNA gene and whole genome shotgun surveys of diverse human samples to identify potential agents and antibiotic resistance profiles. For suspected viral agents we are sequencing randomly primed cDNA libraries and querying these data against a custom viral database. Our results show several new potential bacterial agents that may be associated with different diseases with unknown etiologies. We are currently in the process of translating these methods to the clinic where these approaches can immediately impact human health.

FF0238

**Clade-specific Genomic Signatures as a Method for Accurate Profiling of Metagenomic Datasets**

Patrick Chain

Los Alamos National Lab, Genome Science Group

In the absence of metagenome assembly or when assembly provides a poor picture of the community, characterizing the complexity and relative abundance of microbes in shotgun-sequenced microbiome samples relies on the assignment of individual reads via search of a reference database. There have been efforts to address the complexity of this problem, manipulating either the method we use to search sequences, or modifying the database used for the sequence search. In order to make appropriate and timely conclusions for metagenome datasets, these efforts must also deal with our poor understanding of microbial diversity coupled with the bias in our knowledge of microbial genomes, and the vast amount of data from today's high throughput sequencers. More recent attempts have been made to 'unambiguously' assign reads to a reduced reference database to minimize false positives, however the most successful approaches currently incorporate only those regions originating from gene sequences. In developing a method to identify non-identical regions among a set of sequences, we have created a set of signature sequences based on all bacterial reference genomes, reducing this database by up to 35%, and have tested it on a number of datasets, including human samples spiked with clinically relevant pathogen loads. Using this method, we are able to rapidly obtain taxonomic-specific (Strain, Species, and Genus) profiles, with minimal false positive assignments and compare its utility to other currently available methods.

FF0117

## Metagenomic and Metatranscriptomic Analyses of the Complex Community of a Tropical Wastewater Treatment System

Daniela Drautz3,9, Yehuda Cohen1,2,9, John Gajewski3, Mike Givskov2,4,9, Daniel Huson5,9, Staffan Kjelleberg2,6,9, Peter Little7,9, Rikky Wenang Purbojati9, Hari Seah8, Lynn Tomsho3, Rohan B.H. Williams9, Nicola E. Wittekindt3, Stefan Wuertz9,10,11, Chao Xie9, Fangqing Zhao2,9,12, Stephan C. Schuster2,3,9

1Institute of Life Sciences - Hebrew University of Jerusalem, Jerusalem, Israel 2School of Biological Sciences - Nanyang Technological University, Singapore, Singapore. 3Department of Biochemistry & Molecular Biology - Pennsylvania State University, University Park, United States. 4Department of International Health, Immunology and Microbiology - University of Copenhagen, Copenhagen, Denmark. 5Faculty of Computer Science - University of Tuebingen, Tuebingen, Germany. 6School of Biotechnology and Biomolecular Sciences and Centre for Marine Bio-Innovation - The University of New South Wales, Sydney, Australia. 7Department of Biochemistry - National University of Singapore, Singapore, Singapore 8Public Utilities Board Singapore, Singapore, Singapore. 9Singapore Centre on Environmental Life Sciences Engineering, Singapore, Singapore. 10School of Civil and Environmental Engineering - Nanyang Technological University, Singapore, Singapore.11Department of Civil and Environmental Engineering - University of California Davis, Davis, United States. 12Beijing Institutes of Life Science - Chinese Academy of Sciences, Beijing, China

The establishment of wastewater treatment systems has significantly contributed to improving human health while reducing disease-related mortality. However, exponential human growth over the past few decades has also brought new challenges, such as polluted fresh and marine water environments as a result of wastewater discharge into these systems. These new challenges, together with enhanced nutrient and water recovery from used water, and higher energy efficiency require a better understanding of underlying biological processes of wastewater treatment systems. Using an ultra-deep sequencing approach at a similar scale to that of the Human Microbiome Project, ~12 billion sequence reads generated on the Illumina HiSeq2000 and the Roche/454 FLX platforms were used to investigate the taxonomic as well as functional background of a mostly uncharacterized, extremely diverse floccular microbial community of a tropical wastewater treatment system. To date, this study represents one of the deepest sequencing efforts performed in an environmental context. In total, the collected dataset saturated short-read genomic DNA to 95.2% and long-read genomic DNA to 68%. Metatranscriptomic sequencing of total RNA yielded in 100% sequence saturation of ribosomal RNA whereas the metatranscriptome itself was sequenced to 88% saturation. The metagenomic and metatranscriptomic analyses of this wastewater treatment system have revealed a surprisingly diverse and novel microbial community, comprising more than 2000 genera. However, less than 1% of the sequence data can be aligned to available reference sequences and 37 of the 50 most abundant operational taxonomic units (OTUs) cannot be classified at the genus level. Moreover, high activity in the community is not necessarily associated with the most abundant community members. An aerobic ammonia-oxidizer, for example, only ranked 539th in terms of its abundance within the community but was one of the most active OTUs on the basis of transcription. The complete genomic and transcriptomic dataset from this study provides for a better understanding of the complex biological processes of wastewater treatment systems. Enhanced understanding of these processes will allow for the development of novel engineering bioprocesses which in return will lead to improved ecosystem and urban sustainability.

FF0051b

**CLC bio Products and Supported Applications**

Marta Matvienko

CLC bio

CLC Genomics Workbench enables rapid analysis and visualization of data generated by NGS machines. The user-friendly and intuitive interface allows scientists at all levels of bioinformatics knowledge to analyze the data and search for biological results. Furthermore, the versatile nature of CLC Genomics Workbench allows it to blend seamlessly into existing sequencing analysis workflows, easing implementation and maximizing return on investment. We also present here the recently developed tools, such as Extract Consensus, Transcript Discovery, and Workflows. Blast2GO package was also integrated into CLC genomics Workbench as a plugin.

FF0075

**The CLC Microbial Genome Finishing Module**

Martin Simonsen[1], Marta Matvienko[2], Poul Liboriussen[1], Peder Roed Lindholm Nielsen[1], Jesper Jakobsen[1], Steffen Mikkelsen[1], Henrik Sandmann[1], Søren Mønsted[1], Jannick Dyrløv Bendtsen[1]

[1]CLC bio, Denmark, [2]CLC bio, USA

With the rapid and continuous improvements in sequencing technology it has become possible to produce high quality de novo assemblies of bacterial genomes fast and at a low cost. However, due to e.g. repetitive regions and sequencing errors it is rarely possible to assemble reads into a single error free contig. Consequently, the finishing step, where mis-assemblies are resolved and genomes are closed, has become a bottleneck in genome sequencing.

To make the finishing process more streamlined and hereby reduce the amount of time needed to finishing bacterial genomes, we have developed a plugin for the CLC Genomic Workbench called the CLC Microbial Genome Finishing Module which is a collection of tools for identifying, visualizing and solving problems in genome assemblies. The plugin is fully integrated with the CLC Genomic Workbench which allows e.g. de novo assembly, read mapping and finishing to be performed without the need for importing and exporting data. The tools in the Finishing Module are designed to work with both Sanger and NGS data and they will run efficiently on standard laptop hardware.

Some of the key tools in the Finishing Module include:

- A tool for identifying and annotating mis-assemblies in contigs.
- A contig alignment tool which can align contigs to a reference (or to another set of contigs), visualize alignments and it also provides an intuitive GUI interface for joining, splitting and manually editing contigs.
- An automated tool for fast and easy creation of primers sequences.

The module contains an additional seven tools for genome finishing which can be combined in numerous ways to achieve the desired result. Version 1.0 of The CLC Microbial Genome Finishing Module was released the March 7th and is available for download on www.clcbio.com.

FF0030

**Latest Advances in Bioinformatics Computing**

George Vacek
Director, Life Sciences
Convey Computer Corporation, Richardson, TX, USA
gvacek@conveycomputer.com

Advances in sequencing technology have significantly increased data generation, requiring similar computational advances for bioinformatics analysis. Advanced architectures based on reconfigurable computing can reduce application run times from hours to minutes, while addressing problems unapproachable with commodity servers. The increased capability also improves research quality by allowing more accurate approaches that were previously impractical. This work describes the use of Convey's Hybrid-Core (HC) computing architecture, which combines a traditional x86 system with a reconfigurable coprocessor, to solve data-intensive problems of next-generation sequencing analysis.

Convey developed an alignment kernel that allows HC systems to dramatically reduce time to solution and increase throughput. For example, throughput is improved up to 18x for a full BWA paired-end mapping of human genome sequence data as compared to a traditional server. Additional workflow optimization includes options such as integrated BAM file generation. Extending this to other aspects of variant analysis, key portions of the Genome Analysis Tool Kit (GATK) have been optimized on HC servers. Collaborative efforts with the Broad have dramatically improved the performance of low level kernels such as such as PairHMM and LocusIterator, and thereby higher level applications that call them, like HaplotypeCaller and UnifiedGenotyper. PairHMM is now almost 90x faster on an HC server than it was originally on a traditional server.

A hybrid assembly strategy recently developed by Koren, et al. uses short, high-fidelity sequences to correct the error in long single-molecule sequences. This algorithm for PacBio corrected reads (PBcR) achieves >99.9% base-call accuracy, leading to better assemblies (up to 5x longer median contig sizes) than other sequencing strategies, and even finished genomes. Unfortunately, PBcR requires significant computational run time - in the case of the parrot genome, about 20K core hours – because the all-versus-all overlaps between long- and the short-read sequences use a seed and extend approach based on the Smith-Waterman algorithm. This is an ideal candidate for optimization on an HC system, as Smith-Waterman searches implemented on the coprocessor deliver 15x better throughput than on a conventional server. The overlap subroutine is also a significant step in Celera Assembler's overlap consensus assembler.

FF0149

**Implementing Fast Sequence Analysis Tools Using a Cray XMT2**

Sterling Thomas, Nathan Dellinger, Allan Bolipata, Danielle, Weaver, Tyler Barrus, Daniel Negron, Mitchell Holland

Noblis, Falls Church, Virginia

Advances in sequencing technology have increased the computational demands required for processing, identifying, and analyzing large and complex datasets. Because Single Nucleotide Polymorphisms (SNPs) offer insight into the molecular mechanisms of pathogen evolution, virulence, host preference, lineage calculations, and the emergence of highly pathogenic strains; detecting them rapidly within large sequence readsets would be highly valuable, as well as efficiently performing metagenomic studies of biological communities without prior knowledge of their composition. To address these critical needs, Noblis developed a preliminary suite of novel algorithms that utilizes the strengths of the CRAY XMT 2 supercomputer to perform reference-based multiple sequence alignment (MSA) and SNP detection. This implementation uses next generation sequence reads as input and aligns them against a customized reference library, which can be specific – consisting of one strain or a group of strains belonging to the same species, or highly varied – containing bacterial, viral, and mammalian DNA. The vast resources provided by the CRAY XMT 2 allow MSA and SNP detection at Noblis to proceed at a significantly faster rate than current industry standards without sacrificing precision. As an example, the process of aligning reads and deriving SNPs from 53 million 100 bp reads against a 3,227 Mbp human genome reference sequence was completed in approximately one hour. Additionally, Noblis performed a theoretical metagenomic analysis of 2.7 million sequence reads aligned against a reference library of 16 bacterial genomes with an average size of 5.5 Mbp. This analysis accurately determined the makeup of all microbial species present at a total runtime of 37 seconds. Our results demonstrate rapid sequence alignment of large datasets against a high number of reference genomes, thus creating a method for fast detection of SNPs and accurate identification of all organisms in the metagenomic samples.

FF0221

**Genomics Applications in the Cloud with the DNAnexus Platform**

Andrey Kislyuk, PhD., Sr. Software Engineer

The bioinformatics community faces numerous challenges in developing research and clinical software for analyzing genomic data. The DNAnexus Platform enables solutions to these problems. This presentation will cover its features: massively scalable on-demand cloud infrastructure, fully configurable and scriptable genomics API, command-line interface and SDKs, collaboration and community support, visualization and publication tools, enterprise-grade security, compliance with clinical and diagnostic standards.

We will conduct a brief demo of developing applications on the platform, as well as scientific collaboration, publishing, and reproducibility features.

FF0078

**Sequence Consensus Algorithms and Hierarchical Genome Assembly Process for Effective De Novo Assembly with SMRT® Sequencing**

Aaron Klammer

PacBio

Abstract: The Single Molecule Real-Time (SMRT®) Sequencing platform provides direct sequencing data that can span several thousand bases to tens of thousands of bases in a high-throughput fashion. The capability to get very long reads provides opportunities to get a close-to-finished quality bacteria genome with just a single, long insert (~10 kb or longer) SMRTbell library. Both the necessary read lengths and accuracies for generating good assemblies are accomplished by new algorithmic approaches. The algorithms presented here can construct accurate consensus from the reads where the dominant error modes are insertions and deletions. The long and accurate consensus sequences are used in the following assembly process. We demonstrate how the repeats can be resolved and show the results of such process applied to assemble a bacterial genome and a previously hard-to-assemble Plasmodium falciparum genome having high repeat content and high A/T to G/C ratio.

FF0126

## Haplotype Assembly Refinement & Improvement

Christian Olsen*[1], Kashef Qaadri[1], Matthew Shoa-Azar[2], Joan Wong[2], Trung Nguyen[2], George Rudenko[2], Tina Huynh[2], Aravind Somanchi[2], Jeff Moseley[2], Riyaz Bhat[2], Xinhua Zhao[2,], Scott Franklin[2], Shane Brubaker[2]

1 Biomatters, Inc. 60 Park Place Suite 2100 Newark, NJ 07102

2 Solazyme, Inc. 225 Gateway Blvd. South San Francisco, California 94080

Correspondence: Christian@biomatters.com

Haplotyping, the resolution of two distinct alleles of an organism, is an important and challenging problem in BioInformatics. In particular, the phasing of SNPs (correct assignment of alleles) across long distances is very challenging given today's Next Generation Sequencing (NGS) technologies. We have used Geneious to analyze reads from several technology platforms, including 454, Illumina, and PacBio. Key features of the software including read mapping, read visualization and mate pair analysis, and editing and polymorphism analysis were used to support the workflow. We refined the workflow with an emphasis on speed and accuracy. We have been able to manually haplotype a highly polymorphic, diploid organism over arbitrarily long distances. The process is far faster and more accurate using PacBio reads, although we found a second QC pass with Illumina data to be highly valuable. We verified haplotyping accuracy using known bacterial artificial chromosome (BAC) sequences. In addition, we have explored attempts to perform automated diploid assembly with correct haplotyping phasing, with mixed results. We use the improved Geneious 6.0 *de novo* assembler. We discuss these results here and make suggestions for further improvements to assembly algorithms to help support haplotyped assembly.

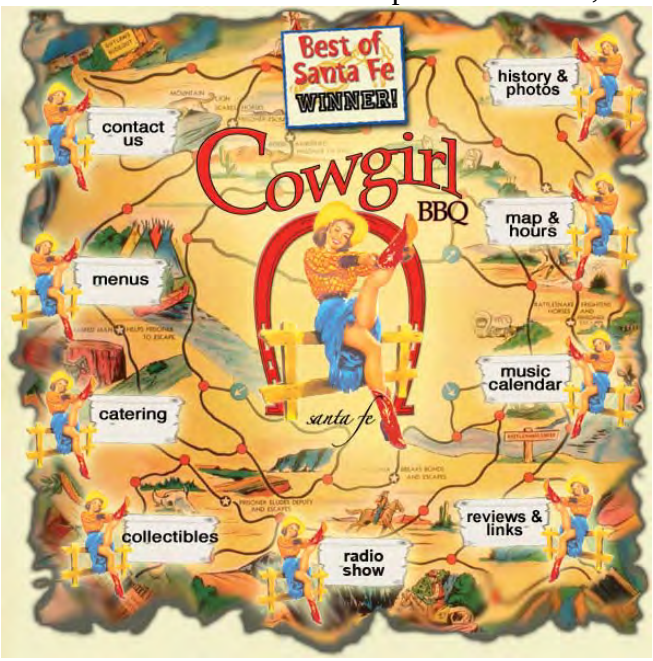
# *Notes*

# *Notes*

# *Happy Hour(s)*

# *Cowgirl BBQ*

505.982.2565  319 S. Guadalupe St  Santa Fe, NM




See map on next page!

5:30pm – 8:00pm, May 30th

Drink tickets (margaritas, beer, sodas) will be provided

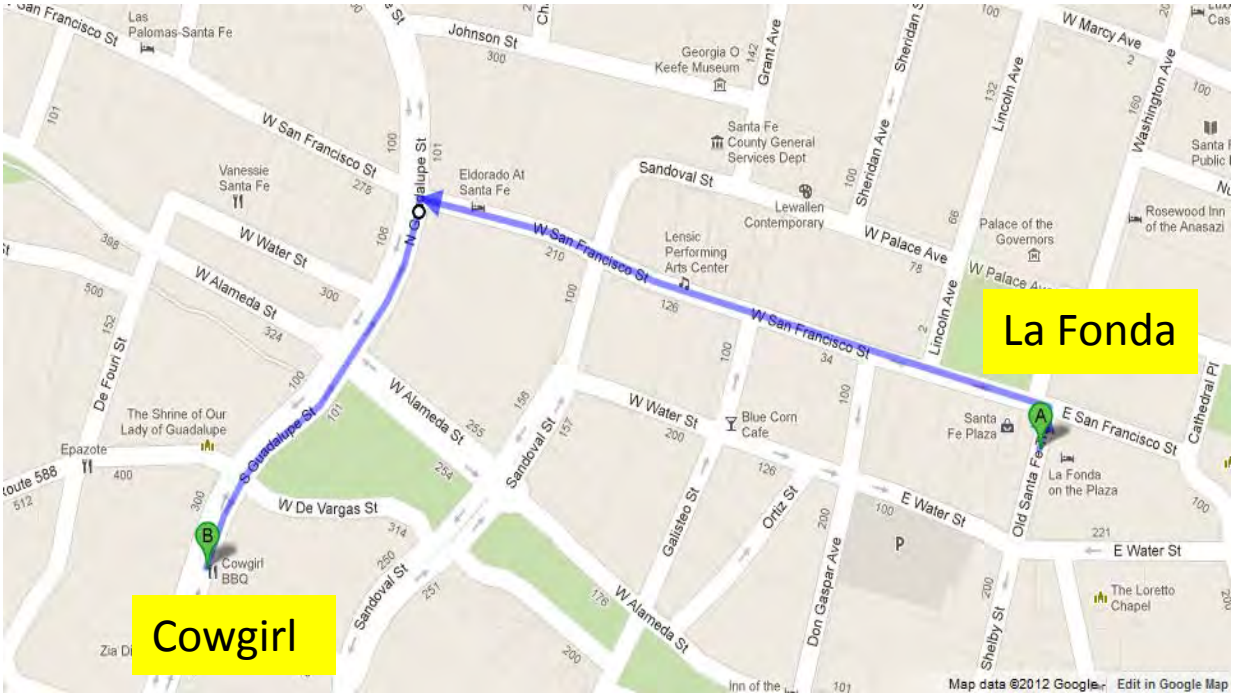
Sponsored by Life Technologies

Enjoy!!!

life technologies™

# *Map to Cowgirl BBQ*

505.982.2565  319 S. Guadalupe St  Santa Fe, NM




# Total Walking Distance

# 0.5 mile, 10 minutes

## The Legend...

Many years ago, when the cattle roamed free and Cowpokes and Cowgirls rode the range, a sassy young Cowgirl figured out that she could have as much fun smokin' meats and baking fine confections as she could bustin' broncs and rounding up outlaws. So she pulled into the fine bustling city of Santa Fe and noticed that nobody in town was making Barbeque the way she learned out on the range. She built herself a Texas-style barbecue pit and soon enough the sweet and pungent scent of mesquite smoke was wafting down Guadalupe street and within no time at all folks from far and near were lining up for heaping portions of tender mesquite-smoked brisket, ribs and chicken. Never one to sit on her laurels, our intrepid Cowgirl figured out that all those folks chowing down on her now-famous BBQ need something to wash it all down with. Remembering a long-forgotten recipe from the fabled beaches of Mexico, she began making the now-legendary Frozen Margarita and the rest, as we say, is History. Before you could say "Tequila!" the musicians were out playing on the Cowgirl Patio and the party was in full swing.

| 05/31/2013 - Friday | | | | |
|---|---|---|---|---|
| **Time** | **Type** | **Abstract #** | **Title** | **Speaker** |
| 7:30 - 8:30am | Breakfast | x | Harvey House Breakfast Buffet | Sponsored by NEB |
| 8:30 - 8:45 | Intro | x | Welcome Back | Chris Detter |
| x | Session Chair | x | Session Chairs | Chair - Patrick Chain<br>Chair - Bob Fulton |
| 8:45 - 9:30 | Keynote | FF0085 | Keep Calm and Carry On as the Human Reference Assembly Updates | Deanna Church |
| 9:30 – 9:50 | Speaker 1 | FF0054 | HAVANA Manual Annotation: the Cartography of a Genome | Mike Kay |
| 9:50 – 10:10 | Speaker 2 | FF0133 | Genomic Analysis of Susceptibility to Autoimmunity | Ward Wakeland |
| 10:10 – 10:30 | Speaker 3 | FF0019 | Cross-platform NGS Method for the Identification STR Loci | Daniel Bornman |
| 10:30 – 10:50 | Speaker 4 | FF0134 | Tomorrow's Genome: Complete Bacterial Genomes in <24 h for Outbreak Response | Ken Dewar |
| 10:50 – 11:10 | Break | x | Beverages and Snacks Provided | Sponsored by AATI |
| 11:10 – 11:30 | Speaker 5 | FF0029 | Consed and BamScape for Next-Gen Sequencing | David Gordon |
| 11:30 – 11:50 | Speaker 6 | FF0137a | Assembling Human Genomes | Jim Knight |
| 11:50 – 12:10 | Speaker 7 | FF0144 | Reducing Assembly Complexity of Microbial Genomes with Single-molecule Sequencing | Adam Phillippy |
| 12:10 - 1:10pm | Lunch | x | Santa Fe Deli Lunch Buffet | Sponsored by illumina |
| x | Session Chair | x | Session Chairs | Chair - Mike Fitzgerald<br>Chair - Darren Grafham |
| 1:10 - 1:30 | Speaker 8 | FF0168 | Reference Assisted Assembly With ALLPATHSLG | Sante Gnerre |
| 1:30 - 1:50 | Speaker 9 | FF0080a | SPAdes: Assembling Microbes in the Cloud | Anton Korobeynikov |
| 1:50 - 2:10 | Speaker 10 | FF0146 | Inexpensive High Quality Genome Assemblies from a Single PCR-free Illumina Library | Ted Sharpe |
| 2:10 - 2:30 | Speaker 11 | FF0097a | Fat-Free Bioinformatics: Successful Microbial Genomics in a Lean Contract Research Environment | Jonathan Jacobs |
| 2:30 - 2:50pm | Closing Discussions | x | Closing Discussions for General Meeting - discuss next year's meeting | Chair - Chris Detter |

# *NOTES*

# *Speaker Presentations (May 31st)*

Abstracts are in order of presentation according to Agenda

Keynote

FF0085

**Keep Calm and Carry On as the Human Reference Assembly Updates**

Deanna M. Church
Staff Scientist, NCBI
Phone 301.594.5695

The human reference assembly (GRCh37) is one of the most important resources in biomedical research today. Producing *de novo* assemblies with current sequencing technology is technically challenging. Alignment of reads to the reference assembly is currently the most robust approach for individual genome analysis today. It is for this reason that we must have the highest quality and most complete reference assembly possible. GRCh37 has been public for four years and several petabytes of sequence data have been aligned to this assembly. Despite known issues and shortcomings of the current reference, the time and computational resources needed to reanalyze this data leads to a sense of trepidation by many when we discuss updating the reference assembly. However, the Genome Reference Consortium (GRC) has released Fix patches for over 100 regions and made 1000s of other, as yet unreleased, fixes to the reference assembly. In some case, these fixes may be quite small; encompassing only a single base, while other fixes involve completely re-tiling misassembled regions. The GRC has released more than 100 of these fixes as assembly patches, but for many, these sequences remain difficult to access and it is clear that integrating these sequences into the full assembly would provide a substantial gain. GRCh38 will be available this fall. During this talk, I will provide examples of updates, why we think these updates are important and tools and strategies for dealing with this transition.

# *NOTES*

FF0054

**HAVANA Manual Annotation: the Cartography of a Genome**

Mike Kay[1], Jonathan Mudge[1], Jennifer Harrow[1]

[1]Wellcome Trust Sanger Institute

The HAVANA group generate manual gene annotation on finished genome sequences. In particular, we aim to describe the complete human transcriptome as part of the GENCODE consortium, as well as all pseudogenes. This means that GENCODE has a higher total transcript count than other public genesets, containing 109,000 models (68%) that are not present in RefSeq or UCSC. This excess is largely explained by our drive to capture all alternatively spliced transcripts and all long non-coding RNAs. However, the incredible output of next-generation RNA sequencing shows that many transcripts remain to be mapped, while a significant proportion of our existing models remain incomplete. HAVANA are also manually annotating the mouse and zebrafish genomes. Furthermore, we are now applying our expertise to the shotgun-based pig and rat genomes in a community-centric manner.

Manual annotation is an excellent way to identify those errors in a genome that affect genes. This includes base errors or indels that disrupt CDS, as well as larger assembly or clone orientation issues that break gene structures. Identified errors are passed onto the Genome Reference Consortium (GRC) for further investigation. If the GRC resolve the issue by creating a fix patch, this sequence is subjected to manual annotation. Furthermore, many assembly issues lie within duplicated regions of the genome, and such regions frequently contain complex gene families of great interest, such as the human leukocyte receptor complex on chromosome 19. The GRC may also create alternative assemblies for such regions, and their annotation is challenging. However, our manual approach is well suited for this task, in particular for distinguishing between coding loci and pseudogenes.

The Wellcome Trust Sanger Institute is operated by Genome Research Limited, a charity registered in England with number 1021457 and a company registered in England with number 2742969, whose registered office is 215 Euston Road, London, NW1 2BE.

FF0133

**Genomic Analysis of Susceptibility to Autoimmunity**

Edward K. Wakeland

University of Texas Southwestern Medical Center

Systemic lupus erythematosus is a complex autoimmune disease characterized by the production of IgG autoantibodies against nuclear antigens and a potent activation of the innate immune system. Genome wide association studies (GWAS) have defined more than 20 risk loci for SLE in Caucasian populations, many of which contain candidate genes with significant regulatory roles in both the adaptive and innate immune systems. The functional variations mediated by these common SLE risk alleles are largely unknown. We have developed an experimental strategy that allows the association of functional information with disease-associated haplotypes at risk loci throughout the genomes of large cohorts of patients and controls. This strategy utilizes targeted resequencing, eQTL datasets, ENCODE annotation information, and the formation of median neighbor joining networks to characterize the functional and disease-association relationships of haplotypes formed by sets of potentially functional variations in tight LD with disease-tagging SNPs. We illustrate this strategy with results for 10 SLE risk loci. We found that risk and resistant haplotypes often differ by multiple functional variations and form highly divergent clades that are associated with either risk or resistance to disease. This analysis further demonstrated that some functional risk haplotypes impact the transcription of multiple genes within or adjacent to the disease associated LD blocks, illustrating the complexity of functional changes associated with common disease risk alleles and demonstrating that a single SLE risk allele can mediate multiple functional changes in the immune system.

FF0019

**Cross-platform NGS Method for the Identification STR Loci**

Daniel Bornman

Identity Management, Battelle Memorial Institute, Columbus, OH 43201

Next generation sequencing (NGS) continues to significantly increase our ability to extract informative details about our genome in the context of not only human health, but also in human identity and characterization. As NGS technology transforms forensic investigations, methods must be established to routinely characterize single nucleotide polymorphisms (SNPs), mitochondrial DNA, and short tandem repeats (STRs) to successfully apply this technology. The later represents the current gold standard marker for identity matching and is currently measured using capillary electrophoresis (CE). In this study, we present a new method for STR typing by targeted sequencing with an analysis of observed sequence error. Data is provided on the typing of 18 STR loci with the identification of partial and variant alleles. We also present a software program (STRTyperTM) developed for generating allele calls from FASTQ files that can be configured as a bioinformatics research tool or as routine forensic genomics laboratory software. A new CMF 3.2-compliant file format (Forensic-DNA Format) for recording and transferring STR sequencing data for forensics use is also presented.

FF0134

**Tomorrow's Genome: Complete Bacterial Genomes in <24 h for Outbreak Response**

Ken Dewar

McGill University, Human Genetics, Montreal, H3A 0G1, Canada; McGill University and Génome Québec Innovation Centre, Montreal, H3A 0G1, Canada

New advances in massively parallel DNA sequencing technologies are promoting a shift away from the push for higher and higher capacity at lower and lower per base costs. Alternative scenarios can now be considered that include very rapid turnaround times for smaller genome projects. Public health and food safety continue to be plagued by bacterial pathogens that can erupt into local or large-scale outbreaks. Rather than focus on a response plan for each pathogen, we are developing a genomics strategy applicable to any bacterial pathogen. Our objective is to progress from raw extracted DNA to a completely sequenced genome and its attendant plasmids, and then to a series of key analyses for typing and trace back, within 24 h of reception of DNA. Our procedure requires no reliance on reference genomes, thus our de novo strategy is capable of detecting a wider range of structural rearrangements, horizontal transfers, extrachromosomal elements, and allelic variations.

The combination of strengths of the MiSeq system for high accuracy basecalling and the PacBio for long read lengths permits the assemblies of very high quality bacterial genomes. It has become standard to obtain completely closed, confirmed circular genomes and plasmids using this approach. The rapid generation of genomes near or exceeding "finished" quality is now a feasible and affordable methodology for deeper investigations into bacterial pathogen evolution and epidemiology.

During early 2013 we undertook a series of blinded tests in collaboration with Canadian health and food safety institutions. Anonymized DNA samples were received and tandem PacBio and MiSeq libraries generated and sequenced. By combining a series of library preparation optimizations and sequencing strategies, we demonstrate we can obtain our first sequencing read datasets by 14.5 h (for preliminary species identification), complete PacBio and MiSeq runs in less than 19 h, and generate combined MiSeq/PacBio assemblies by 21 h. Residual gap closing (using existing data) and project specific analyses (virtual karyotypes, genome and plasmid summaries, PCR primers for isolate specific detection) are reported back to the outbreak investigators within 24 h. The entire process is supported with password protected custom instances of the UCSC genome browser which includes real time updates of sequencing progress and access to all assemblies and annotation results.

FF0029

**Consed and BamScape for Next-Gen Sequencing**

David Gordon and Phil Green, University of Washington, Seattle, Washington, USA

BamScape is a new program that displays an overview of reads in a BAM file using little memory, and can bring up Consed to view and edit targeted regions. It can search for problem regions, defined as high (or low) depth of coverage regions, regions of high rates of inconsistent mate pairs, or regions in which a large fraction of the reads are discrepant with the reference. It can do this interactively or in batch, outputting in Picard format. BamScape plots read depth, depth of reads having inconsistently mapped mates, and read discrepancy rates (including indels).

BamScape will keep track of which regions you have edited in consed so when you are done editing, you can run a program (part of consed) to modify your reference sequence. You must then realign your reads to the reference if you want an updated BAM file.

Consed can now show tracks in the manner of the UCSC Genome Browser.
It can show gene tracks and conservation scores from a BED file, and cam show graphs from a WIG fixedStep file.

When adding new reads to an assembly, an approximate location for each read can be specified. This is useful when there are regions of similar sequence and you know which region that reads should go to.

Support has been added for RNA-Seq alignments.

Reads of any type (e.g., Illumina) can be edited.

The read name highlight feature has been expanded: you can highlight mates of highlighted reads, you can highlight reads from a file, and you can highlight reads by string at the cursor. Highlighted reads can be removed as a group or used for tearing a contig.

These and other features will be discussed.

FF0137a

**Assembling the Human Genome**

Jim Knight

Roche

FF0144

**Reducing Assembly Complexity of Microbial Genomes with Single-molecule Sequencing**

Adam Phillippy1, Sergey Koren1, Gregory Harhay2, Timothy Smith2, James Bono2, Dayna Harhay2, D. Scott McVey3, Diana Radune1, Nicholas Bergman1

1National Biodefense Analysis and Countermeasures Center, Frederick, Maryland 21702
2USDA, ARS, U.S. Meat Animal Research Center, Clay Center, NE 68933
3School of Veterinary Medicine and Biomedical Sciences, University of Nebraska, Lincoln, NE
68583

Genome assembly algorithms cannot fully reconstruct microbial chromosomes from the short DNA reads output by either first or second-generation sequencing. As a result, most genomes are left unfinished due to the significant resources required to manually close gaps left in the draft assemblies. Single-molecule sequencing addresses this problem by greatly increasing sequencing read length, which simplifies the assembly problem. To measure the benefit of single-molecule sequencing on microbial genome assembly, we analyzed the repeat complexity of all 2,265 previously finished bacteria and archaea, and sequenced the genomes of six bacteria using the PacBio RS, Roche 454, and Illumina MiSeq instruments for comparison. Our results suggest that current single-molecule sequencing technology can close more than 70% of known bacterial and archaeal genomes at finished-grade quality using only a single sequencing library. In addition, this single-library approach shows comparable accuracy to hybrid assemblies of multiple technologies. This drastically reduces the cost of microbial finishing to below $2,000 per genome, in most cases, and enables high-fidelity, population-scale studies of pangenomes and chromosomal organization.

# *NOTES*

# Lunch

**12:10 – 1:10pm**

## Sponsored by

# *Notes*

FF0168

**Reference Assisted Assembly With ALLPATHSLG**

Sante Gnerre, Bruce Walker, Sarah Young, Terry Shea, Aaron Berlin, Sakina Saif, Amr Abouelleil, Alma Imamovic, David Jaffe, Chad Nusbaum

Broad Institute of MIT and Harvard, Cambridge MA

ALLPATHSLG has proven to be a highly effective *de novo* genome assembler of short-read sequencing data for a wide range of genomes, from bacteria to mammals. However, as more finished references and high-quality draft genomes become available, there is an increasing opportunity to use the genomes of close relatives to assist the assembly process. This is not a new idea; in particular, we previously developed algorithm extensions to the ARACHNE assembler which successfully enabled the assisted assembly process of mammalian genomes using low-coverage Sanger sequencing [1]. Here, we will present new algorithms which enable ALLPATHSLG to make use of a nearby reference genome when assembling Illumina short-read sequencing data. We limited ourselves to approaches which are “safe” in that they never rely on reference information alone to make connectivity decisions in the genome; all joins must have supporting evidence in the read data.

We will also present results of this technique as applied to a variety of microbial genome assembly projects, where we often see a significant improvement in scaffolding as well as a reduction in the number of contigs in the resulting assemblies by 50% or more.
This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.:HHSN272200900018C
[1] Assisted assembly: how to improve a de novo genome assembly by using related species.

Gnerre S, Lander ES, LindbladToh K, Jaffe DB. Genome Biol.
2009;10(8):R88. doi: 10.1186/gb2009108r88.
Epub 2009 Aug 27.

FF0080a

**SPAdes: Assembling Microbes in the Cloud**

Anton Korobeynikov[1], Alexei Gurevich[1], Pavel Pevzner[1,2], Alla Lapidus[1]

[1] St. Petersburg Academic University, St. Petersburg, Russia
[2] University of California, San Diego, USA

SPAdes – St. Petersburg genome assembler - was designed to assemble specific sequencing data produced for single-cell microbial projects. At the time there were no other assembly tools capable of dealing with MDA data challenges (highly non uniform read coverage, an elevated number of chimeric reads and chimeric read pairs, as well as an elevated number of sequencing errors). Currently there is a good number of assemblers in use and researches are faced with the task of having to choose which one is best for the particular experiments they are planning. To help benchmark assemblers and assess the quality of the resulting assemblies another tool called QUAST was developed in our LAB. This tool is very easy to use and provides detailed analysis of assemblies. To quickly disseminate both good quality open source genomic tools to researchers all over the world, we have chosen to use the cloud-based DNAnexus platform. The functionality of the platform allows for a seamless creation of the entire data-processing pipeline starting from the reads to assembly and quality assessment, ending with annotations, and other sophisticated analyses.

This work was supported by the Government of the Russian Federation (grant 11.G34.31.0018).

FF0146

**Inexpensive High Quality Genome Assemblies from a Single PCR-free Illumina Library**

Ted Sharpe, Shuangye Yin, Neil Weisenfeld, Bayo Lau, Louise Williams, Diana Tabbaa, Andi Gnirke, Ryan Hegarty, Carsten Russ, Chad Nusbaum, Iain MacCallum, David B Jaffe

Broad Institute, Genome Sequencing and Analysis Program, Cambridge, MA, 02142

The ability to rapidly, accurately and inexpensively determine microbial genome sequences remains a critical goal for biology and medicine. One approach is to mix data types, for example 100 base paired reads from a fragment library and a jumping library, or even better, those together with very long PacBio reads. However the cost of these approaches is proportionally high. Using multiple data types also increases variability and delivery time.

We propose an approach which starts from a single fragment library, which is made without PCR, thus reducing bias. From this library, paired 250 base reads are then generated either by Illumina MiSeq or HiSeq 2500, depending on the number of samples that can be pooled. This single data type turns out to be extraordinarily powerful. While this data type cannot disambiguate perfect repeats longer than ~500 bases, the assemblies can be otherwise nearly perfect, and their overall quality can easily exceed that from multiple library approaches.

Next these data are assembled using our new method, DISCOVAR. Briefly, (1) a new approach to error correction which involves finding all the 'friends' of a given read; (2) generation of a graph using fixed K; (3) graph simplification via local increase of K. The output of the assembly is a graph that can be represented essentially verbatim or encapsulated in the assembly representation format FASTG. Where a reference sequence from another strain is available, the output also includes a catalog of differences between the two strains.

We piloted this approach on the GC-rich bacterium Rhodobacter spheroides. Comparison to the finished reference sequence reveals that the entire genome (two chromosomes plus five plasmids) are present without gaps, with ~1 error per Mb. As a graph, the assembly also reveals valid features not present in the reference sequence, but present in a subpopulation of the sample.

FF0097a

**Fat-Free Bioinformatics: Successful Microbial Genomics in a Lean Contract Research Environment**

Jonathan Jacobs [1]*, Richard Winegar [2]

* Cooresponding Author
[1] MRIGlobal, Rockville, MD
[2] MRIGlobal, Palm Bay, FL

MRIGlobal supports a variety of industrial and government clients with a broad range of programs that include basic research, clinical studies, and operational support services. As a contract research organization, we are often faced with responding to our client's needs in real-time, with a high priority on executing programs efficiently and on-budget. As demands for next generation sequencing continues to increase, so too does the demand for flexible bioinformatics analysis pipelines for a diverse set of applications. For example, current microbial genomics programs at MRIGlobal include: analyses of viral quasispecies, host-pathogen interactions, genome engineering, environmental surveillance, and whole genome sequencing. To meet these challenges, our next-generation sequencing and bioinformatics programs are lean, task-oriented, and focused on program deliverables. In this seminar, we present an overview of how we continue to meet these challenges using a combination of open-source and commercial bioinformatics tools such as CLC Genomics Workbench.

# *Notes*

# *Notes*

# 2013 SFAF Attendee List

| FF # | Name | Affiliation | email |
|---|---|---|---|
| FF0001 | Chris Detter | Los Alamos National Laboratory (LANL) | cdetter@lanl.gov |
| FF0002 | David Bruce | Los Alamos National Laboratory (LANL) | dbruce@lanl.gov |
| FF0003 | Olga Chertkov | Los Alamos National Laboratory (LANL) | ochrtkv@lanl.gov |
| FF0004 | Hajni Daligault | Los Alamos National Laboratory (LANL) | hajkis@lanl.gov |
| FF0005 | Ashlynn Daughton | Los Alamos National Laboratory (LANL) | adaughton@lanl.gov |
| FF0006 | Tracy Erkkila | Los Alamos National Laboratory (LANL) | terkkila@lanl.gov |
| FF0007 | Cheryl Gleasner | Los Alamos National Laboratory (LANL) | cdgle@lanl.gov |
| FF0008 | Lynne Goodwin | Los Alamos National Laboratory (LANL) | lynneg@lanl.gov |
| FF0009 | Wei Gu | Los Alamos National Laboratory (LANL) | wgu@lanl.gov |
| FF0010 | Adam Kotorashvili | Richard G. Lugar Center for Public Health Research, Georgia | adam.kotorashvili@gmail.com |
| FF0011 | Nadia Fedorova | J. Craig Venter Institute (JCVI) | nfedorov@jcvi.org |
| FF0012 | Alex Hutcheson | Pacific Biosciences | ahutcheson@pacificbiosciences.com |
| FF0013 | Alexander Kozik | Genome Center, University of California, Davis | akozik@atgc.org |
| FF0014 | Alfredo Lopez De Leon | Novozymes | ALLO@novozymes.com |
| FF0015 | Roman Aranda | Defense Forensics Science Center | roman.aranda3.ctr@mail.mil |
| FF0016 | Robert J Baker | Texas Tech University | Robert.Baker@ttu.edu |
| FF0017 | Beth Nelson | Novozymes | BANE@novozymes.com |
| FF0018 | Rachel Bartholomew | Pacific Northwest National Lab | rachel.bartholomew@pnnl.gov |
| FF0019 | Daniel Bornman | Battelle Memorial Institute | bornmand@battelle.org |
| FF0020 | Cliff Han | Los Alamos National Laboratory (LANL) | han_cliff@lanl.gov |
| FF0021 | Shannon Johnson | Los Alamos National Laboratory (LANL) | shannonj@lanl.gov |
| FF0022 | Yuliya Kunde | Los Alamos National Laboratory (LANL) | y.a.kunde@lanl.gov |
| FF0023 | Daniela Campanella | J. Craig Venter Institute (JCVI) | dcampane@jcvi.org |
| FF0024 | Cecilie Boysen | CLC Bio | cboysen@clcbio.com |
| FF0025 | Trevor Wagner | OpGen | twagner@opgen.com |
| FF0026 | Jonathon Foos | OpGen | jfoos@opgen.com |
| FF0027 | Chad Locklear | Integrated DNA Technologies | clocklear@idtdna.com |
| FF0028 | Daniel Bozinov | Genimbi | dbozinov@genimbi.com |
| FF0029 | David Gordon | University of Washington | dgordon@u.washington.edu |
| FF0030 | George Vacek | Convey Computer Corporation | gvacek@conveycomputer.com |
| FF0031 | Harper VanSteenhouse | BioNano Genomics | hvansteenhouse@bionanogenomics.com |
| FF0032 | Heng Dai | Bionano Genomics | hdai@bionanogenomics.com |
| FF0033 | Jane Hutchinson | Roche Diagnostics | jane.hutchinson@roche.com |
| FF0034 | Jodi Irwin | Federal Bureau of Investigation | Jodi.Irwin@ic.fbi.gov |
| FF0035 | Jason Farlow | University Partnership Program | jasonfarlow@hotmail.com |
| FF0036 | Ric Sugarek | Integrated DNA Technologies | rsugarek@idtdna.com |
| FF0037 | John Gillece | TGen North | jgillece@tgen.org |
| FF0038 | Nathan Hicks | TGen North | nhicks@tgen.org |
| FF0039 | Jo Wood | Wellcome Trust Sanger Institute | jmdw@sanger.ac.uk |
| FF0040 | Guy Griffiths | Wellcome Trust Sanger Institute | gg3@sanger.ac.uk |
| FF0041 | John Havens | Integrated DNA Technologies | jhavens@idtdna.com |
| FF0042 | Todd Dickinson | BioNano Genomics | tdickinson@bionanogenomics.com |
| FF0043 | Xing Yang | BioNano Genomics | xyang@bionanogenomics.com |
| FF0044 | Ernest Lam | BioNano Genomics | elam@bionanogenomics.com |
| FF0045 | Jason J. LeBlanc | Defense Forensic Science Center | jason.j.leblanc9.ctr@mail.mil |
| FF0046 | Lijing Bu | University of New Mexico | lijing@unm.edu |
| FF0047 | Jon Longmire | Los Alamos National Laboratory (LANL) | jonlongmire@lanl.gov |
| FF0048 | Lori Peterson | Caldera Pharmaceuticals | peterson@cpsci.com |
| FF0049 | Nicole Touchet | Caldera Pharmaceuticals | nltouchet@yahoo.com |
| FF0050 | Luke Hickey | Pacific Biosciences | lhickey@pacificbiosciences.com |
| FF0051 | Marta Matvienko | CLC Bio | mmatvienko@clcbio.com |
| FF0052 | Matt Dunn | Sanger Institute | md3@sanger.ac.uk |
| FF0053 | Michael FitzGerald | Broad Institute | fitz@broadinstitute.org |
| FF0054 | Mike Kay | Wellcome Trust Sanger Institute | mpk@sanger.ac.uk |

| | | | |
|---|---|---|---|
| FF0055 | Teri Rambo Mueller | Roche Applied Science | teri.mueller@roche.com |
| FF0056 | Owatha (Tootie) Tatum | LBL-JGI | oltatum@lbl.gov |
| FF0057 | Caleb Phillips | Texas Tech University | caleb.phillips@ttu.edu |
| FF0058 | Ron Walters | Pacific Northwest Nat'l Laboratory | ron@ron-walters.com |
| FF0059 | Ravi Kumar | Novozymes | RVKU@novozymes.com |
| FF0060 | Sarah Buddenborg | University of New Mexico | sbuddenb@unm.edu |
| FF0061 | Sarah Young | The Broad Institute | stowey@broadinstitute.org |
| FF0062 | Scott P Layne | Alfred E Mann Foundation | scottl@aemf.org |
| FF0063 | Scott Rose | Integrated DNA Technologies | srose@idtdna.com |
| FF0064 | Todd Smith | PerkinElmer | Todd.Smith@PERKINELMER.COM |
| FF0065 | Fiona Stewart | New England Biolabs | stewart@neb.com |
| FF0066 | Cheryl L. Tarr | Centers for Disease Control and Prevention (CDC) | crt6@cdc.gov |
| FF0067 | David Trees | Centers for Disease Control and Prevention (CDC) | dlt1@cdc.gov |
| FF0068 | Eija Trees | Centers for Disease Control and Prevention (CDC) | eih9@cdc.gov |
| FF0069 | Angie Trujillo | Centers for Disease Control and Prevention (CDC) | awt0@cdc.gov |
| FF0070 | Maryann Turnsek | Centers for Disease Control and Prevention (CDC) | hud4@cdc.gov |
| FF0071 | Edward Wack | MIT Lincoln Laboratory | wack@ll.mit.edu |
| FF0072 | Warren Andrews | BioNano Genomics | wandrews@bionanogenomics.com |
| FF0073 | Mary Ann Allen | OpGen | mallen@opgen.com |
| FF0074 | Yilin Zhang | Elim Biopharmaceuticals | yilin@elimbio.com |
| FF0075 | Martin Simonsen | CLC Bio | msimonsen@clcbio.com |
| FF0076 | Sandra Porter | Digital World Biology | sandra@digitalworldbiology.com |
| FF0077 | Po-E Li | Los Alamos National Laboratory (LANL) | po-e@lanl.gov |
| FF0078 | Aaron Klammer | Pacific Biosciences | aklammer@pacificbiosciences.com |
| FF0079 | Chien-Chi Lo | Los Alamos National Laboratory (LANL) | chienchi@lanl.gov |
| FF0080 | Anton Korobeynikov | Saint Petersburg State University, Russia | anton@korobeynikov.info |
| FF0081 | Christian Buhay | Baylor College of Medicine - HGSC | cbuhay@bcm.edu |
| FF0082 | Sophie Mangenot-Layac | CEA-Genoscope | mangenot@genoscope.cns.fr |
| FF0083 | Karine Labadie | CEA-Genoscope | klabadie@genoscope.cns.fr |
| FF0084 | Stefen Engelen | CEA-Genoscope | sengelen@genoscope.cns.fr |
| FF0085 | Deanna Church | NCBI | church@ncbi.nlm.nih.gov |
| FF0086 | David W. Cleary | Dstl Porton Down | DWCLEARY@mail.dstl.gov.uk |
| FF0087 | Phil A. Rachwal | Dstl Porton Down | PARACHWAL@dstl.gov.uk |
| FF0088 | Cyrille Longin | CEA-Genoscope | clongin@genoscope.cns.fr |
| FF0089 | Darren Grafham | Sheffield Children's Hospital | darrengrafham@gmail.com |
| FF0090 | Gary L. Simpson | University of New Mexico | garyl.simpson@comcast.net |
| FF0091 | George Rosenberg | University of New Mexico | ghrose@unm.edu |
| FF0092 | George VanDegrift | Convey Computer | gvandegrift@conveycomputer.com |
| FF0093 | Richard Gibbs | Baylor College of Medicine | agibbs@bcm.edu |
| FF0094 | Harold Lee | Pacific Biosciences | hlee@pacificbiosciences.com |
| FF0095 | J. Enrique Herrera-Galeano | NMRC-Frederick | Jesus.Herrera.ctr@med.navy.mil |
| FF0096 | Howard Cash | Gene Codes Corporation | howardc@genecodes.com |
| FF0097 | Jonathan Jacobs | MRIGlobal | jjacobs@mriglobal.org |
| FF0098 | Kim McMurry | Los Alamos National Laboratory (LANL) | kmcmurry@lanl.gov |
| FF0099 | John S. Oliver | Nabsys | oliver@nabsys.com |
| FF0100 | Judy Ney | KAPA Biosystems | judy.le@kapabiosystems.com |
| FF0101 | Maryke Appel | KAPA Biosystems | maryke.appel@kapabiosystems.com |
| FF0102 | Julien Tremblay | Department of Energy (DOE) | jtremblay@lbl.gov |
| FF0103 | Phil Latreille | Monsanto Company | phil.latreille@monsanto.com |
| FF0104 | Brad Langhorst | New England Biolabs | |
| FF0105 | Jing Lu | Monsanto Company | jing.lu@monsanto.com |
| FF0106 | Marta Orlikowska | University of Texas Southwestern Medical Center | Marta.Orlikowska@UTSouthwestern.edu |
| FF0107 | Lee Ann McCue | Pacific Northwest National Lab | leeann.mccue@pnnl.gov |
| FF0108 | Ginger Metcalf | Baylor College of Medicine - HGSC | metcalf@bcm.edu |
| FF0109 | Michael Rey | Novozymes | MWR@novozymes.com |
| FF0110 | Raquel Bromberg | University of Texas Southwestern Medical Center | raquel.bromberg@utsouthwestern.edu |

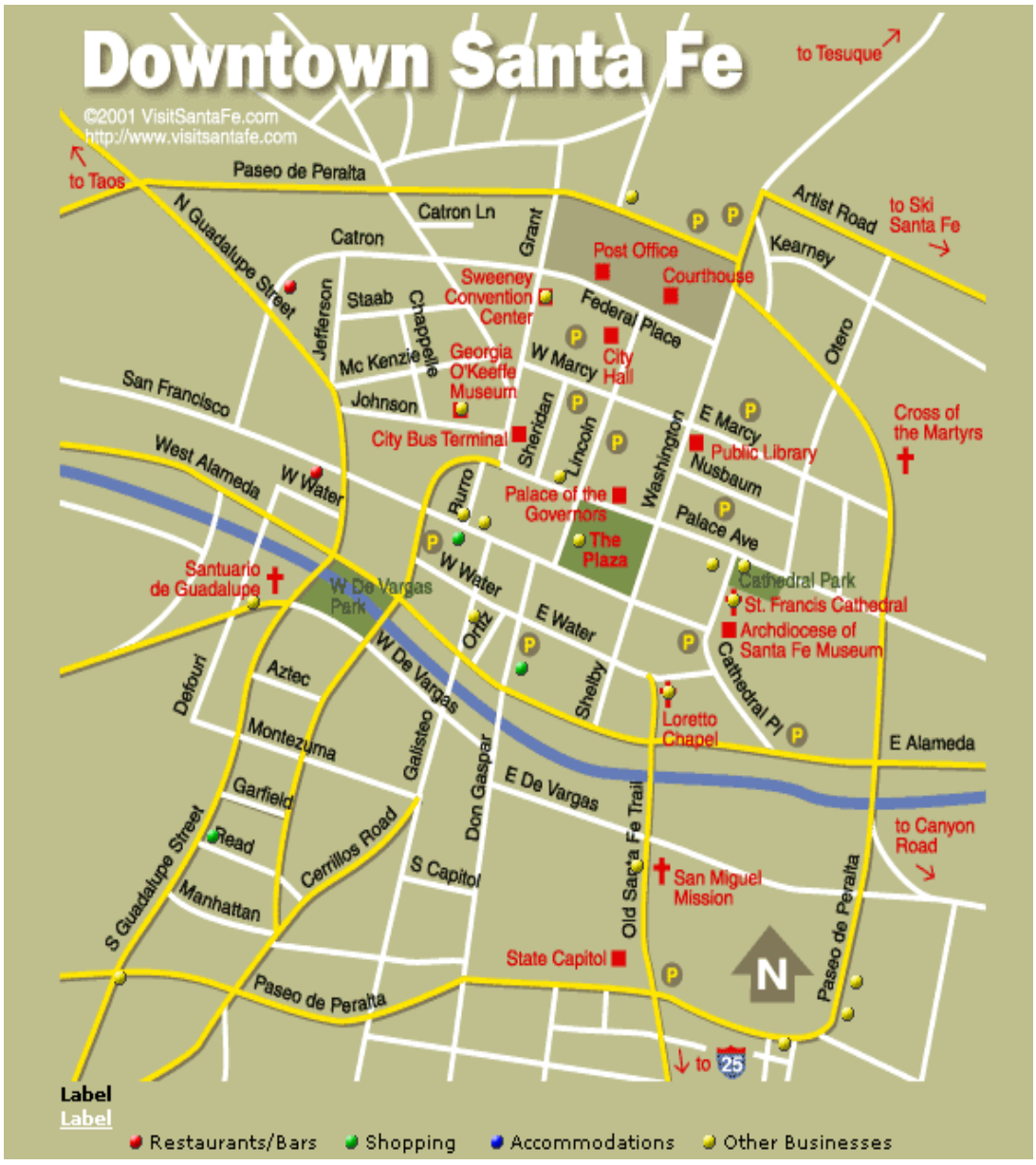
| | | | |
|---|---|---|---|
| FF0111 | Tod P. Stuber | U.S. Dept. of Agriculture | Tod.P.Stuber@aphis.usda.gov |
| FF0112 | Scott Sammons | Centers for Disease Control and Prevention (CDC) | zno6@cdc.gov |
| FF0113 | Beverly Parson-Quintana | Los Alamos National Laboratory (LANL) | bapq@lanl.gov |
| FF0114 | Louis Sherman | Purdue University | lsherman@purdue.edu |
| FF0115 | Catherine Smith | Centers for Disease Control and Prevention (CDC) | cab2@cdc.gov |
| FF0116 | Stacie Smith | Monsanto Company | stacie.t.smith@monsanto.com |
| FF0117 | Stephan Schuster | Pennsylvania State University | scs@bx.psu.edu |
| FF0118 | Krista Reitenga | Los Alamos National Laboratory (LANL) | reitenga@lanl.gov |
| FF0119 | Timothy Sussman | Los Alamos National Laboratory (LANL) | tsussman@lanl.gov |
| FF0120 | Shawn Starkenburg | Los Alamos National Laboratory (LANL) | shawns@lanl.gov |
| FF0121 | Mark D. Adams | J. Craig Venter Institute (JCVI) | madams@jcvi.org |
| FF0122 | Anitha Sundararajan | National Center for Genome Resources (NCGR) | asundara@ncgr.org |
| FF0123 | Asif Khalak | Pacific Biosciences | akhalak@pacificbiosciences.com |
| FF0124 | Bud Mishra | New York University and Cold Spring Harbor Lab | bud.mishra@gmail.com |
| FF0125 | Caitlin Wolf | University of New Mexico | cwolf92@unm.edu |
| FF0126 | Christian Olsen | Biomatters | christian@biomatters.com |
| FF0127 | Cristina Takacs-Vesbach | University of New Mexico | cvesbach@unm.edu |
| FF0128 | Yadhu Kumar | GATC Biotech AG, Germany | kumar@gatc-biotech.com |
| FF0129 | Faye Schilkey | National Center for Genome Resources (NCGR) | fds@ncgr.org |
| FF0130 | Vibeke Halkjaer-Knudsen | Sandia National Laboratories | vnhalkj@sandia.gov |
| FF0131 | Bill Arndt | Sandia National Laboratories | wdarndt@sandia.gov |
| FF0132 | Jerry W. Dragoo | University of New Mexico | jdragoo@unm.edu |
| FF0133 | Ward Wakeland | UT Southwestern Medical Center | Edward.Wakeland@UTSouthwestern.edu |
| FF0134 | Ken Dewar | McGill University | ken.dewar@mcgill.ca |
| FF0135 | Robert Fulton | The Genome Institute/Washington University School of Medicine | bfulton@genome.wustl.edu |
| FF0136 | Tina Lindsay | The Genome Institute/Washington University School of Medicine | tgraves@genome.wustl.edu |
| FF0137 | James Knight | Roche | james.knight@roche.com |
| FF0138 | Kristen Knipe | Centers for Disease Control and Prevention (CDC) | wgg9@cdc.gov |
| FF0139 | Stanley Langevin | Sandia National Labs | salange@sandia.gov |
| FF0140 | Mindy Luce | SeqWright | Mindy.T.Luce@ge.com |
| FF0141 | Milind Misra | University of New Mexico | misra@unm.edu |
| FF0142 | Neha Varghese | Joint Genome Institute (JGI) | njvarghese@lbl.gov |
| FF0143 | TBD | TBD…do not remove | |
| FF0144 | Adam Phillippy | National Biodefense Analysis and Countermeasures Center | phillippya@nbacc.net |
| FF0145 | Surya Saha | Cornell University | ss2489@cornell.edu |
| FF0146 | Ted Sharpe | The Broad Institute | tsharpe@broadinstitute.org |
| FF0147 | Terrance Shea | The Broad Institute | tshea@broadinstitute.org |
| FF0148 | Thiru Ramaraj | National Center for Genome Resources (NCGR) | tr@ncgr.org |
| FF0149 | Sterling Thomas | Noblis | Sterling.Thomas@noblis.org |
| FF0150 | Hazuki Teshima | Los Alamos National Laboratory (LANL) | hazuki@lanl.gov |
| FF0151 | Xiaoben Jiang | University of New Mexico | sdpapet@gmail.com |
| FF0152 | Helen Cui | Los Alamos National Laboratory (LANL) | hhcui@lanl.gov |
| FF0153 | Dan Colman | University of New Mexico | dcolman@unm.edu |
| FF0154 | Darren Lee | Nabsys | darren@nabsys.com |
| FF0155 | David J. Van Horn | University of New Mexico | davevanh@gmail.com |
| FF0156 | David Alexander | Pacific Biosciences | dalexander@pacificbiosciences.com |
| FF0157 | Elliott Franco Drábek | University of Maryland - IGS | elliott.drabek@gmail.com |
| FF0158 | Adam English | Baylor College of Medicine | english@bcm.edu |
| FF0159 | Heather Buelow | University of New Mexico | hnbuelow@gmail.com |
| FF0160 | Joseph F. Petrosino | Baylor College of Medicine | jpetrosi@bcm.edu |
| FF0161 | Krista Ternus | Signature Science | kternus@signaturescience.com |
| FF0162 | Kelsy Thompson | Oklahoma State University | kelsyrt@ostatemail.okstate.edu |
| FF0163 | Kate Weinbrecht | Oklahoma State University | kateldw@okstate.edu |
| FF0164 | Ken Taylor | Advanced Analytical | KTaylor@aati-us.com |
| FF0165 | Keven Stevens | Integrated DNA Technologies | kstevens@idtdna.com |
| FF0166 | Norman Doggett | Los Alamos National Laboratory (LANL) | doggett@lanl.gov |

| | | | |
|---|---|---|---|
| FF0167 | Shannon Dugan | Baylor College of Medicine | sdugan@bcm.edu |
| FF0168 | Sante Gnerre | The Broad Institute | sante@broadinstitute.org |
| FF0169 | Joann Mudge | National Center for Genome Resources (NCGR) | jm@ncgr.org |
| FF0170 | Gerwald Koehler | Oklahoma State University | gerwald.kohler@okstate.edu |
| FF0171 | Maqsudul Alam | University of Hawaii | maqsudul@gmail.com |
| FF0172 | Nito Panganiban | Tulane National Primate Research Center | apangani@tulane.edu |
| FF0173 | Shanavaz Nasarabadi | IntegenX | ShanavazN@IntegenX.com |
| FF0174 | Steve Siembieda | Advanced Analytical | SSiembieda@aati-us.com |
| FF0175 | Marine Murtskhvaladze | Ilia State University, Georgia | dna_lab@iliauni.edu.ge |
| FF0176 | Gvantsa Chanturia | National Center for Disease Control and Public Health, Georgia | romail28@gmail.com |
| FF0177 | Tea Tevdoradze | National Center for Disease Control and Public Health, Georgia | t_tevdoradze@yahoo.com |
| FF0178 | Maka Kokhreidze | Laboratory of the Ministry of Agriculture, Georgia | makakokhreidze@yahoo.com |
| FF0179 | Omayma Al-Awar | Illumina | oalawar@illumina.com |
| FF0180 | Andrew Farmer | National Center for Genome Resources (NCGR) | adf@ncgr.org |
| FF0181 | Ben Mans | Agricultural Research Council of South Africa | MansB@arc.agric.za |
| FF0182 | Bob Dietrich | Syngenta Biotechnology | bob.dietrich@syngenta.com |
| FF0183 | Cathy Cleland | Los Alamos National Laboratory (LANL) | ccleland@lanl.gov |
| FF0184 | Darrell Dinwiddie | University of New Mexico | dldinwiddie@cmh.edu |
| FF0185 | Charmaine D'Souza | Life Technologies | Charmaine.D'Souza@lifetech.com |
| FF0186 | Lori M. Gladney | Centers for Disease Control and Prevention (CDC) | hze1@cdc.gov |
| FF0187 | TBD | TBD…do not remove | |
| FF0188 | Fernanda Henry | Oklahoma State University | fernanda.henry@okstate.edu |
| FF0189 | Kelly Hoon | Life Technologies (Ion Torrent) | hoon@lifetech.com |
| FF0190 | David Jasper Gilbert Rees | Agricultural Research Council of South Africa | reesj@arc.agric.za |
| FF0191 | Jochen Kumm | ILRI, Kenya | jochen.kumm@gmail.com |
| FF0192 | Johar Ali | Ontario Institute for Cancer Research | Johar.Ali@oicr.on.ca |
| FF0193 | Jeremy Ledermann | Centers for Disease Control and Prevention (CDC) | bpj7@cdc.gov |
| FF0194 | George Michuki | ILRI, Kenya | G.Michuki@cgiar.org |
| FF0195 | Vakhtang BERISHVILI | Richard G. Lugar Center for Public Health Research, Georgia | vberishvili@gmail.com |
| FF0196 | Brigita KARDAVA | CH2M HILL | brigita.kardava@ch2m.com |
| FF0197 | Shanmuga Sozhamannan | Critical Reagent Repository (CRP) | shanmuga.sozhamannan.ctr@mail.mil |
| FF0198 | Malin Young | Sandia National Laboratories | mmyoung@sandia.gov |
| FF0199 | Matt Salter | Noblis | Matt.Salter@noblis.org |
| FF0200 | Nate Dellinger | Noblis | Nathan.Dellinger@noblis.org |
| FF0201 | Sagar Utturkar | University of Tennessee | sutturka@utk.edu |
| FF0202 | Zedias Chikwambi | Agricultural Research Council of South Africa | ChikwambiZ@arc.agric.za |
| FF0203 | Hazem Iskandar Sari Haddad | Arramtha / Jordan University of Science and Technology | sjaradat@just.edu.jo |
| FF0204 | Areej Mohammad Qar'an | Arramtha / Jordan University of Science and Technology | areejalquran@yahoo.com |
| FF0205 | Bart Weimer | School of Veterinary Medicine, University of California, Davis | bcweimer@ucdavis.edu |
| FF0206 | Steve Turner | Pacific Biosciences | sturner@pacificbiosciences.com |
| FF0207 | Cody Cain | Illumina | ccain@illumina.com |
| FF0208 | Callum J. Bell | National Center for Genome Resources (NCGR) | cjb@ncgr.org |
| FF0209 | Archana Chauhan | The University of Tennessee | achauha1@utk.edu |
| FF0210 | Glenn Tesler | University of California, San Diego | gptesler@math.ucsd.edu |
| FF0211 | Jason Smith | Pacific Biosciences | jrsmith@pacificbiosciences.com |
| FF0212 | Daniel Mazur | Life Technologies | Daniel.Mazur@lifetech.com |
| FF0213 | TBD | TBD..do not remove | |
| FF0214 | Joshua L. Santarpia | Sandia National Laboratories | jsantar@sandia.gov |
| FF0215 | Jeff Kaszak | Illumina | jkaszak@illumina.com |
| FF0216 | Masoud Toloue | Bioo Scientific | mtoloue@biooscientific.com |
| FF0217 | Peter Ngam | National Center for Genome Resources | pbn@ncgr.org |
| FF0218 | Linda Ray | Beckman Coulter, Inc. | lray@beckman.com |
| FF0219 | Graham Threadgill | Beckman Coulter Life Sciences | gthreadgill@beckman.com |
| FF0220 | Andreas Sundquist | DNAnexus | andreas@dnanexus.com |
| FF0221 | Andrey Kislyuk | DNAnexus | akislyuk@dnanexus.com |
| FF0222 | Prachi Nakashe | Bioo Scientific | prachiN@biooscientific.com |

| | | | |
|---|---|---|---|
| FF0223 | Jonathan Bingham | Google | binghamj@google.com |
| FF0224 | Donna Muzny | Baylor College of Medicine | donnam@bcm.edu |
| FF0225 | Phelix Majiwa | Agricultural Research Council-Onderstepoort Veterinary Institute, Pretoria, South Africa | MajiwaP@arc.agric.za |
| FF0226 | Zakee Sabree | The Ohio State University | sabree.8@osu.edu |
| FF0227 | Clotilde Teiling | Illumina, Inc. | cteiling@illumina.com |
| FF0228 | Tyler Thacker | USDA-Agricultural Research Service | tyler.thacker@ars.usda.gov |
| FF0229 | Jason Affourtit | Life Technologies | Jason.affourtit@lifetech.com |
| FF0230 | Razvan Mathias | Google | raz@google.com |
| FF0231 | Christian Whitchurch | DTRA | christian.whitchurch@dtra.mil |
| FF0232 | Heley Fiske | Illumina, Inc. | hfiske@illumina.com |
| FF0233 | Gang Xie | Los Alamos National Laboratory (LANL) | xie@lanl.gov |
| FF0234 | Karen Davenport | Los Alamos National Laboratory (LANL) | kwdavenport@lanl.gov |
| FF0235 | Armand Dishcosa | Los Alamos National Laboratory (LANL) | armand@lanl.gov |
| FF0236 | Bin Hu | Los Alamos National Laboratory (LANL) | binhu@lanl.gov |
| FF0237 | Ben McMahon | Los Alamos National Laboratory (LANL) | mcmahon@lanl.gov |
| FF0238 | Patrick Chain | Los Alamos National Laboratory (LANL) | pchain@lanl.gov |
| FF0239 | Sanaa Ahmed | Los Alamos National Laboratory (LANL) | sahmed@lanl.gov |
| FF0240 | Hong Shen | Los Alamos National Laboratory (LANL) | xshen@lanl.gov |
| FF0241 | Tracey Freitas | Los Alamos National Laboratory (LANL) | traceyf@lanl.gov |
| FF0242 | Momo Vuyisich | Los Alamos National Laboratory (LANL) | vuyisich@lanl.gov |
| FF0243 | Shihai Feng | Los Alamos National Laboratory (LANL) | sfeng@lanl.gov |
| FF0244 | Matthew Scholz | Los Alamos National Laboratory (LANL) | mscholz@lanl.gov |
| FF0245 | Frank Boellmann | Illumina, Inc. | fboellmann@illumina.com |
| FF0246 | Kashef Qaadri | Biomatters, Inc. | kashef@biomatters.com |
| FF0247 | Jim George | Illumina, Inc. | jgeorge@illumina.com |
| FF0248 | Hemant Mohapatra | Google | hmohapatra@google.com |
| FF0249 | Chris Rampey | IDT | jheddens@idtdna.com |
| FF0250 | Isaac Meek | PerkinElmer | Isaac.Meek@PERKINELMER.COM |
| FF0251 | Terrill Yazzie | University of New Mexico | tyazzi01@gmail.com |
| FF0252 | Shelley MacNeil | University of New Mexico | smacneil88@gmail.com |
| FF0253 | Rougeron Virginie | Medical Research Center, Franceville, GABON | rougeron.virginie@gmail.com |
| FF0254 | Chris Hopkins | Centers for Disease Control & Prevention | vqd8@cdc.gov |
| FF0255 | Yan Xu | Los Alamos National Laboratory (LANL) | yxu@lanl.gov |
| FF0256 | Ahmet Zeytun | Los Alamos National Laboratory (LANL) | azeytun@lanl.gov |
| FF0257 | Lucy Zhang | Los Alamos National Laboratory (LANL) | xlz@lanl.gov |

# *NOTES*

# *Map of Santa Fe, NM*



# *History of Santa Fe, NM*

Thirteen years before Plymouth Colony was settled by the Mayflower Pilgrims, Santa Fe, New Mexico, was established with a small cluster of European type dwellings. It would soon become the seat of power for the Spanish Empire north of the Rio Grande. Santa Fe is the oldest capital city in North America and the oldest European community west of the Mississippi.

While Santa Fe was inhabited on a very small scale in 1607, it was truly settled by the conquistador Don Pedro de Peralta in 1609-1610. Santa Fe is the site of both the oldest public building in America, the Palace of the Governors and the nation's oldest community celebration, the Santa Fe Fiesta, established in 1712 to commemorate the Spanish reconquest of New Mexico in the summer of 1692. Peralta and his men laid out the plan for Santa Fe at the base of the Sangre de Cristo Mountains on the site of the ancient Pueblo Indian ruin of Kaupoge, or "place of shell beads near the water."

The city has been the capital for the Spanish "Kingdom of New Mexico," the Mexican province of Nuevo Mejico, the American territory of New Mexico (which contained what is today Arizona and New Mexico) and since 1912 the state of New Mexico. Santa Fe, in fact, was the first foreign capital over taken by the United States, when in 1846 General Stephen Watts Kearny captured it during the Mexican-American War.

Santa Fe's history may be divided into six periods:

**Preconquest and Founding**
**(circa 1050 to 1607)**

Santa Fe's site was originally occupied by a number of Pueblo Indian villages with founding dates from between 1050 to 1150. Most archaeologists agree that these sites were abandoned 200 years before the Spanish arrived. There is little evidence of their remains in Santa Fe today.

The "Kingdom of New Mexico" was first claimed for the Spanish Crown by the conquistador Don Francisco Vasques de Coronado in 1540, 67 years before the founding of Santa Fe. Coronado and his men also discovered the Grand Canyon and the Great Plains on their New Mexico expedition.

Don Juan de Onate became the first Governor-General of New Mexico and established his capital in 1598 at San Juan Pueblo, 25 miles north of Santa Fe. When Onate retired, Don Pedro de Peralta was appointed Governor-General in 1609. One year later, he had moved the capital to present day Santa Fe.

**Settlement Revolt & Reconquest**
**(1607 to 1692)**

For a period of 70 years beginning the early 17th century, Spanish soldiers and officials, as well as Franciscan missionaries, sought to subjugate and convert the Pueblo Indians of the region. The indigenous population at the time was close to 100,000 people, who spoke nine basic languages and lived in an estimated 70 multi-storied adobe towns (pueblos), many of which exist today. In 1680, Pueblo Indians revolted against the estimated 2,500 Spanish colonists in New Mexico, killing 400 of them and driving the rest back into Mexico. The conquering Pueblos sacked Santa Fe and burned most of the buildings, except the Palace of the Governors. Pueblo Indians occupied Santa Fe until 1692, when Don Diego de Vargas reconquered the region and entered the capital city after a bloodless siege.

**Established Spanish Empire**
**(1692 to 1821)**

Santa Fe grew and prospered as a city. Spanish authorities and missionaries - under pressure from constant raids by nomadic Indians and often bloody wars with the Comanches, Apaches and Navajos-formed an alliance with Pueblo Indians and maintained a successful religious and civil policy of peaceful coexistence. The Spanish policy of closed empire also heavily influenced the lives of most Santa Feans during these years as trade was restricted to Americans, British and French.

**The Mexican Period**
**(1821 to 1846)**

When Mexico gained its independence from Spain, Santa Fe became the capital of the province of New Mexico. The Spanish policy of closed empire ended, and American trappers and traders moved into the region. William Becknell opened the l,000-mile-long Santa Fe Trail, leaving from Arrow Rock, Missouri, with 21 men and a pack train of goods. In those days, aggressive Yankeetraders used Santa Fe's Plaza as a stock corral. Americans found Santa Fe and New Mexico not as exotic as they'd thought. One traveler called the region the "Siberia of the Mexican Republic."

For a brief period in 1837, northern New Mexico farmers rebelled against Mexican rule, killed the provincial governor in what has been called the Chimayó Rebellion (named after a village north of Santa Fe) and occupied the capital. The insurrectionists were soon defeated, however, and three years later, Santa Fe was peaceful enough to see the first planting of cottonwood trees around the Plaza.

**Territorial Period**
**(1846 to 1912)**

On August 18, 1846, in the early period of the Mexican American War, an American army general, Stephen Watts Kearny, took Santa Fe and raised the American flag over the Plaza. Two years later, Mexico signed the Treaty of Guadalupe Hidalgo, ceding New Mexico and California to the United States.

In 1851, Jean B. Lamy, arrived in Santa Fe. Eighteen years later, he began construction of the Saint Francis Cathedral. Archbishop Lamy is the model for the leading character in Willa Cather's book, "Death Comes for the Archbishop."

For a few days in March 1863, the Confederate flag of General Henry Sibley flew over Santa Fe, until he was defeated by Union troops. With the arrival of the telegraph in 1868 and the coming of the Atchison, Topeka and the Santa Fe Railroad in 1880, Santa Fe and New Mexico underwent an economic revolution. Corruption in government, however, accompanied the growth, and President Rutherford B. Hayes appointed Lew Wallace as a territorial governor to "clean up New Mexico." Wallace did such a good job that Billy the Kid threatened to come up to Santa Fe and kill him. Thankfully, Billy failed and Wallace went on to finish his novel, "Ben Hur," while territorial Governor.

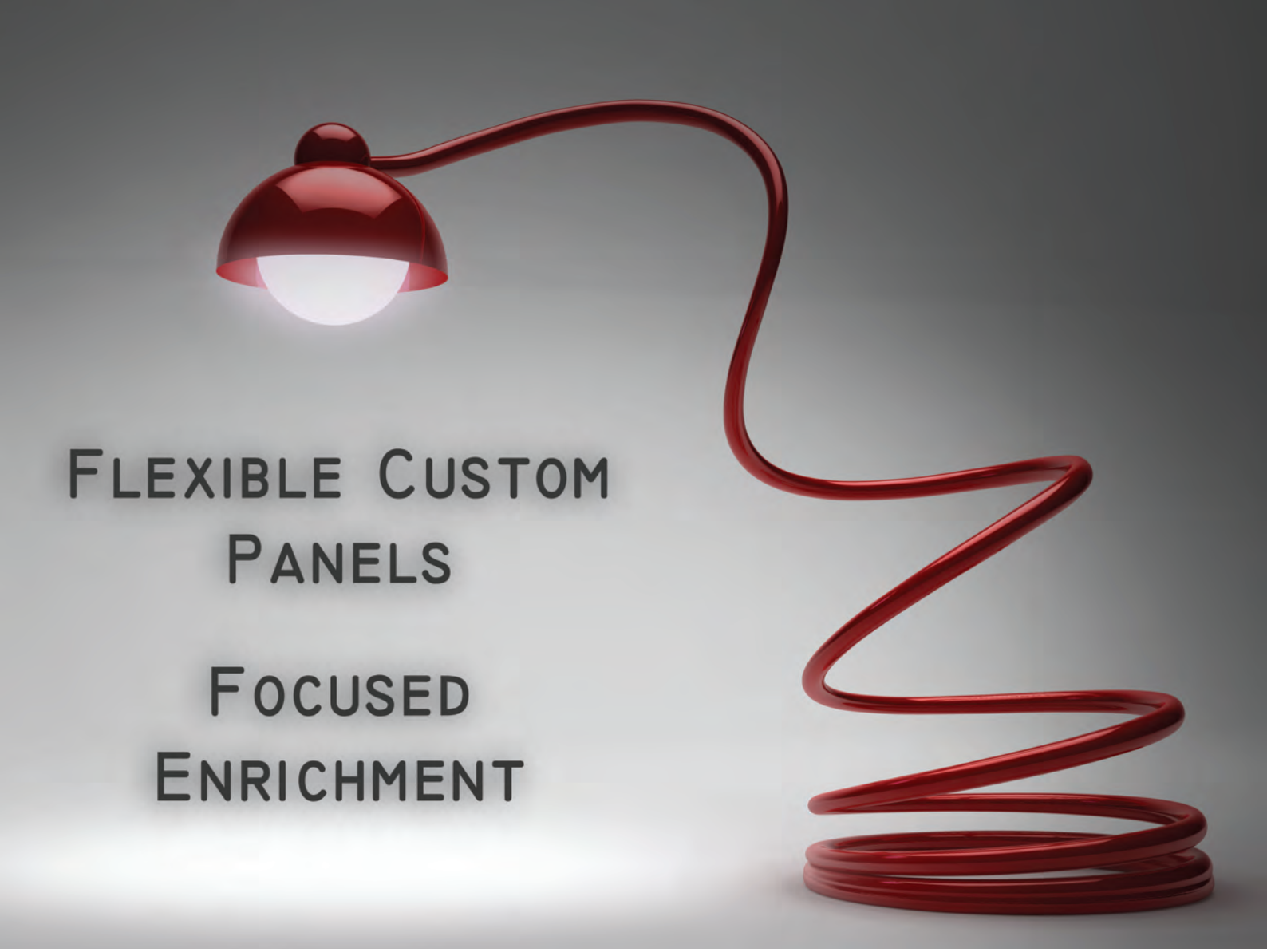
**Statehood**
**(1912 to present)**

When New Mexico gained statehood in 1912, many people were drawn to Santa Fe's dry climate as a cure for tuberculosis. The Museum of New Mexico had opened in 1909, and by 1917, its Museum of Fine Arts was built. The state museum's emphasis on local history and native culture did much to reinforce Santa Fe's image as an "exotic" city.

Throughout Santa Fe's long and varied history of conquest and frontier violence, the town has also been the region's seat of culture and civilization. Inhabitants have left a legacy of architecture and city planning that today makes Santa Fe the most significant historic city in the American West.

In 1926, the Old Santa Fe Association was established, in the words of its bylaws, "to preserve and maintain the ancient landmarks, historical structures and traditions of Old Santa Fe, to guide its growth and development in such a way as to sacrifice as little as possible of that unique charm born of age, tradition and environment, which are the priceless assets and heritage of Old Santa Fe."

Today, Santa Fe is recognized as one of the most intriguing urban environments in the nation, due largely to the city's preservation of historic buildings and a modern zoning code, passed in 1958, that mandates the city's distinctive Spanish-Pueblo style of architecture, based on the adobe (mud and straw) and wood construction of the past. Also preserved are the traditions of the city's rich cultural heritage which helps make Santa Fe one of the country's most diverse and fascinating places to visit.




# Target Enrichment for Next Generation Sequencing

## xGen™ Lockdown™ Probes for Target Capture

xGen™ Lockdown™ Probes are individually synthesized oligos for next generation sequencing target enrichment and custom panel development. Each probe is assessed by mass spectrometry for quality control, enabling a seamless transition from discovery to clinical application.

- Start your project faster with a 7–10 day turnaround time.
- Develop custom panels by pooling up to 2000 probes per tube.
- Transition to high volume applications easily with 3 production scales
- Cost-effective target capture with a low price per enrichment

IDT®
INTEGRATED DNA TECHNOLOGIES

THE CUSTOM BIOLOGY COMPANY

WWW.IDTDNA.COM

100%

# *"Sponsors"*


http://www.pacificbiosciences.com/
Primary Meeting Sponsor


http://www.lifetechnologies.com
Cowgirl Happy Hour


http://www.roche-diagnostics.us/
Meet and Greet Party


http://www.illumina.com/
Lunch and Keynote

# *"Sponsors"*

NEW ENGLAND BioLabs Inc.

http://www.neb.com
Fruit and Juice - Breakfast

IDT INTEGRATED DNA TECHNOLOGIES

http://www.idtdna.com/
Meeting Guides

PerkinElmer®
*For the Better*

http://www.perkinelmer.com/
Lunch and Keynote

ADVANCED ANALYTICAL

http://www.aati-us.com/
Break and Keynote

# *"Sponsors"*


http://www.kapabiosystems.com/company
Break


http://www.opgen.com/
Break


http://www.novozymes.com
Break


https://dnanexus.com/
Break


http://www.bionanogenomics.com/
Break


http://www.clcbio.com/
Break

# *"Sponsors"*









Thank You !!!



NEW ENGLAND BioLabs Inc.